NATURALISING INTENTIONALITY: STRUCTURAL ISOMORPHISM AND THE DETERMINATION OF MENTAL CONTENT

Craig Files

Department of Philosophy
University of Adelaide

Submitted January 2000 for the degree of Doctor of Philosophy

For the opportunity to pursue my chosen path, and the respect, support, and encouragement given to my choice, I dedicate this work to my father.

CONTENTS

ABSTRACT

A *resemblance* theory of the intentionality or contents of mental states is almost universally assumed a non-starter by contemporary philosophers. My overall aim in this work is to show that a revised version of the resemblance theory termed the *structural isomorphism* theory of content determination (TCD) presents a viable – indeed, highly competitive – alternative to the orthodox theories. The structural isomorphism TCD operates with a *second-order* isomorphism: the values of a represented feature are systematically reflected by variations in a non-equivalent physical feature of a representing system. A physically structured neural representing system represents a variable by virtue of mirroring its *structure*. I begin with an overview of representationalism, and how two traditional problems with the resemblance theory can be overcome. The work then divides into three main, overlapping parts. First, I pit the structural isomorphism TCD against the "orthodox approach" to a particularly recalcitrant problem, that of accommodating *misrepresentation*. I show how the orthodox TsCD fail to resolve the problem of misrepresentation, and develop an alternative approach with structural isomorphism at its heart, one which provides a natural and principled means of accommodating misrepresentation. The middle part of the thesis investigates the notion of second-order isomorphism at length, weds the structural isomorphism theory to a "microcontent" conception of mental content, and defends structural isomorphism against the charge of content indeterminacy. Two domains of representational activity – temperature and colour – are employed to argue against the standard externalist account of content determination, and for a significant re-conception of mental content in terms of a *notional* world, a world *as-represented*. In the final part I investigate the prevailing assumption that content determination is orthogonal to cognitive science. I show that while *classicism* is not committed to a particular TCD, the orthogonality assumption is *false* of *connectionism*. Structural isomorphism is the mechanism which "disciplines" the processing of representation bearers in connectionist systems. Structural isomorphism, that is, renders connectionism a *computational* framework, and so connectionism is committed to

the structural isomorphism TCD. Overall, structural isomorphism articulates a perspective on content determination which accords a crucial role to the physical properties of mental representation bearers themselves.

STATEMENT OF ORIGINALITY

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

17/08/00

ACKNOWLEDGEMENTS

My greatest intellectual debt is to Gerard O'Brien; a three-way thank-you. For the insightful comments on earlier drafts of this work. For the many stimulating discussions we had over the years on a number of the issues discussed herein. And for the inspiring and penetrating lectures I was fortunate to experience as an undergraduate. There started what this work makes a start on.

I would also like to acknowledge the valuable forum for learning and exchange of ideas provided by the 'Cognitive Science/Philosophy of Mind Reading Group' at the University of Adelaide. Thanks to the participants, in the group's various incarnations: Diarmuid Crowley, Steve Crowley, Greg Currie, Denise Gamble, Peter Lavskis, Gerard O'Brien, Greg O'Hair, Jon Opie, Belinda Paterson, Vladimir Popescu, and Ian Ravenscroft.

To those who helped in a different, equally invaluable manner, I am grateful to my parents for encouraging the kinds of attitudes in which I find so much satisfaction. I thank my father for always understanding just what this work meant to me. Thanks to my friends, my brother Ashley included, for (and at their considerable peril) gentle and genuine inquiries. Particular mention goes to my friend Bruce Dwyer who also provided vital computer assistance (dragging me into the 1980's just before the century ends). And special thanks to Briar, for her loving patience and support, and for my life outside philosophy.

Finally, and on a more formal note, I'd like to acknowledge the following two factors.

Part of this work was undertaken while I held an Australian Postgraduate Research Award. I am grateful for the financial assistance provided by such a scheme, administered through the Commonwealth Government of Australia's Department of Employment, Education, Training, and Youth Affairs.

The line of argument pursued in subsection 3.2., Chapter One, is based on my 'Goodman's Rejection of Resemblance', published in the *British Journal of Aesthetics*, October 1996, pp. 398-412.

PREFACE

It rained for weeks and weeks and we were all *so* tired of Pin the Symbol on the Horse and the Intentional Prance, but not even Uncle Churchland could remember how to play Eliminativism. Some of the children were getting cranky and started throwing BB-pellets at the frog (who, admittedly, didn't seem to mind what it snapped its tongue at). Others drifted into the corner to dust off the old Consciousness puzzles, and could be heard muttering determinedly "We're going to take this game seriously *this* time." It was going to be a terribly *boring* afternoon until seemingly out of nowhere appeared Old Uncle Hume's Imitation Game, which no one had seen in *ages*. Some of the older children soon began to snigger and scoff behind their hands ("Such a silly game", they hissed, "it never works"), and tried to snatch it away. But some of the younger children looked interested. Sure the instructions were long gone, and it was so worn and faded that some of the pieces didn't *exactly* match. But perhaps, they thought, it could be repaired. And at least it was something *different* to play with. So soon everyone was grabbing and poking and talking at once and the whole room became *very* lively. But Auntie said there needed to be some shush and someone would have an eye out with that thing if we didn't please settle down a bit.

(With apologies to Jerry Fodor 1985/1990a, 3).

INTRODUCTION

REVIVING RESEMBLANCE

Since the demise of behaviourism, philosophy of mind has generally come to embrace that which the folk have known all along: mental states have legitimate ontological status in their own right as mediators between environmental stimuli and behavioural responses. Moreover, mental states operate as such internal intermediaries by virtue of a remarkable property they possess. They are, according to the folk, *information-bearing* items. Mental states carry information about a cognitive agent's environment, as well as information concerning the agent's own psychological and bodily condition. Or, as philosophers of mind like to say, mental states have the property of *intentionality* or "aboutness". They represent (stand-in for, connect with, refer to, are directed-at, or point to) features, objects, and states of affairs beyond themselves. That is, as it is variously expressed in the philosophical literature, mental states possess representational content, meaning, or semantic value. And as the folk well know, the intentionality of the mental has a crucial role in the causal explanation of behaviour. We behave in the way we do because of the particular representational contents of our mental states. According to contemporary philosophy of mind, the mind is a *representational* system, and it is the representational properties of mental states which drive our behaviour. With the importance of intentionality to a proper understanding of mind fully recognised, the task becomes one of explaining how such a property is generated. It is to such a task that the present work is directed.

While the study of intentionality has a long history in philosophy, contemporary interest is usually traced to Franz Brentano (1874/1973). In one sense, contemporary philosophers can come close to agreeing with Brentano's oft-cited remark that intentionality is "the mark of the mental". With the rise of modern cognitivism, intentionality assumed centre-stage in scientific thinking about the mind, and the prevailing view is that intentionality and consciousness are the two defining features of the mental, properties which set mental phenomena apart from all other phenomena. Representation of the external environment by various types of sensory states dominates current philosophical attention.[1] In addition, certain higher-order mental states, typically beliefs and desires (but also memories and intentions), are cited as clear examples of content-bearing internal entities, and again, often it is the state of the external world that such thoughts represent. The immediate environment, however, is not the only domain of representational activity. Mental states are often about distant (both spatially and temporally) states of affairs, and even non-existent or fictional entities and scenarios. It is also common to acknowledge that a cognitive agent's mental states can be about other mental states the agent possesses, as well as certain bodily conditions.

In another sense, however, the vast majority of contemporary philosophers of mind disagree with Brentano's claim that intentionality is "the mark of the mental". For what Brentano meant by his remark is that the intentionality of the mental serves to distinguish mental phenomena from *physical* phenomena. Contemporary philosophy of mind, however, adopts a thoroughly naturalistic framework. For one thing, this involves a commitment to materialism. Mental phenomena, that is, are purely physical entities which arise from the physical activity of the brain and some kind of naturalistically specifiable relationship between brain and environment. Consequently, the task these days is to explain how a physical world gives rise to physical phenomena which exhibit the remarkable property of "aboutness". For another thing, not only is intentionality assumed to arise as a result of purely physical processes, it is to be explained using the resources of the natural sciences, in terms that are themselves non-intentional or non-semantic. To do otherwise would render an explanation circular. Intentionality is not an ultimate and irreducible feature of the world. As Jerry Fodor (1987, 97) has quipped, "if intentionality is real, it must be really something else." What is required, then, is a reductive explanation of intentionality or mental representation. Such a project is the *naturalisation* of intentionality, an undertaking which has been one of the predominant research industries of philosophy of mind for pretty much the past two decades. It is to such a project that the present work aims to contribute.

---

[1] Some philosophers draw a distinction between the 'intentionality' of certain folk psychological states and the 'representational' properties of perceptual and sensory states. This is one of the few distinctions in the realm of mental representation that I won't be concerned to enforce in this work, essentially because I don't think it's particularly robust or principled. Perceptions and sensations, to my mind, are as much as part of folk psychological explanation as beliefs and desires (though their operation is typically tacit).

The contemporary practice is to see the task of explaining the intentionality of the mental in terms of an answer to a more particular question: what is it by virtue of that mental states possess representational *content*? Saying that mental states possess "aboutness" means, of course, that any given mental state is about something in particular. What, specifically, a mental state represents is its representational content. The question, then, is how is mental content conferred or determined? This issue – the "*question of content determination*", as I'll call it – will be the specific topic with which the present work is concerned. What I'll be offering is a "*theory of content determination*" (TCD). [2]

The theory I'm selling here is a "new" approach to the question of content determination. It is one which has received very little in the way of attention and development, and one which differs significantly from the orthodox proposals. It is, however, a version of an old idea, namely, the *resemblance* theory. What I'm going to develop in this work is a revised or updated resemblance theory known as the *structural isomorphism* theory of mental content determination.[3] In outline, the structural isomorphism TCD is a version of resemblance theory which results from relaxing the requirement that the resemblance relation be understood as a *first-order* isomorphism. A first-order isomorphism, the type of isomorphism operative in the traditional resemblance theory, is one where the representing items, or as I'll say, "representation bearers", have physical properties in common with the items being represented, the "representational objects".[4] The structural isomorphism TCD, in contrast, operates with a *second-order* understanding of the resemblance relation. Representation bearers do not share physical properties with their representational objects, rather a system of representation bearers mirrors the *structure* of a system of representational objects. That is, variation among a system

[2] The term "theory of content determination" is due to Von Eckardt (1993). Often philosophers talk of a "theory of representation", but for reasons which will emerge below (Chp. One, sec. 2.), I think this terminology can be misleading.

[3] In describing the structural isomorphism TCD above as a "new" approach to the question of content determination, I put 'new' in scare quotes for two reasons. First, as indicated, it is a variation on the traditional resemblance theory. Second, a handful of theorists have put forward a structural isomorphism theory, though not always under such a description. See Gardenfors (1996), Palmer (1978), Swoyer (1991), and Von Eckardt (1993). The theory has received its most detailed defence in Robert Cummins' (1996) *Representations, Targets, and Attitudes*. Cummins calls it the "picture" theory. I therefore cannot lay claim to pioneering the development of the structural isomorphism TCD.

[4] 'Representation' is ambiguous between the item doing the represent*ing* and the item being represent*ed*, hence the distinction, using terminology due to Von Eckardt (1993), between the representation bearer and the representational object. Some writers have recently begun to talk of the "representational vehicle" in referring to what I call the representation bearer.

of representational objects is tracked, or mirrored, by corresponding variation in a physically non-equivalent property of the representing system.

The mere mention of the resemblance theory is enough to send many contemporary philosophers of mind into howls of protest. While the question of content determination strongly divides the philosophy of mind community, there is one thing upon which the overwhelming majority agrees: the resemblance theory is not a live option. My overall aim in this work is to show that such an attitude is mistaken. Over the following pages I intend to revive the resemblance theory, and, in the modified form of structural isomorphism, demonstrate that it is a serious contender in the quest to satisfactorily answer the question of content determination.

But why a "new" approach to the question of content determination? After all, contemporary philosophy of mind is certainly not bereft of ideas. In recent times there has emerged a number of candidate answers: the causal theory, most prominently developed by Jerry Fodor (1987, 1990a, 1994); the biological function approach, or teleosemantics, pioneered by Ruth Millikan (1984, 1993); the hybrid causal-biological function account defended in the work of Fred Dretske (1986/1990a, 1988, 1995); the neo-behaviourist "intentional stance" position championed by Daniel Dennett (1987); and a more or less loose cluster of ideas – which, in general terms can be characterised as (causal-role) functionalism applied to the question of content determination – in the form of functional role semantics and conceptual role semantics (see, for example, Block 1986, 1987; Harman 1982). I'll have cause to examine each of these theories below. The general point at this stage, however, is that contemporary philosophy of mind is a long way from a consensus as regards the best prospect for satisfactorily resolving the issue of content determination. All of the existing theories, as is well documented, suffer from various significant difficulties, such that, it's fair to say, we are some way short of having a satisfactory solution to the problem of mental content determination. My perspective is that this unsatisfactory situation is unlikely to improve without a rather significant change in thinking. The naturalisation of intentionality project, to put it bluntly, has become just a little stagnant. Perhaps none of the existing theories is on the right track. A large part of the motivation behind the present work is that it's time for a fresh approach to the determination of mental content.

Why, then, the *resemblance* theory? Why seek to resuscitate a theory widely thought to be fundamentally unsuited to the task of explaining mental content? In broad terms, the rationale is that given the current state of the naturalisation of intentionality project, it is important that no possible avenue to a satisfactory explanation of mental content be ruled-out. And the resemblance theory is one area of the content-determination landscape which has remained largely unexplored. More particularly, and in a more "combative" vein, my belief that the resemblance theory deserves a detailed investigation derives from two quarters: one, a much-discussed problem in the philosophy of mind; the other, certain considerations from within cognitive science. I'll take these in turn.

The one problem which, more than any other, plagues the naturalisation of intentionality project is that of *misrepresentation*. The fact is that mental states can be in error; they can misrepresent. A TCD that is in any way remotely plausible must allow for the capacity for misrepresentation. But the problem of *accommodating* representational error has proven particularly recalcitrant. Though the issue has dominated critical discussion in the mental representation literature, existing "orthodox" proposals to allow for misrepresentation have not exactly met with widespread enthusiasm.[5] Indeed, the lack of a viable answer to the problem of misrepresentation is perhaps the main source of disaffection with the TsCD currently on offer. In this work, rather than survey all the difficulties faced by existing TsCD, I will focus on the critical issue of accommodating misrepresentation. And my approach to this problem will be pretty much the reverse of the prevailing methodology. Rather than formulating a TCD and then attempting to let-in misrepresentation, as is the usual strategy, I am going to *begin* with misrepresentation. That is, my strategy will be to build-in the capacity for representational error from the outset, such that, in a sense, the problem of misrepresentation drives an answer to the question of content determination. The TCD arrived at by taking this route, I will show below, is structural isomorphism.[6]

[5] By "orthodox" TsCD I mean those theories other than structural isomorphism. Sometimes I'll just speak of "current" TsCD or "existing" TsCD with no disrespect intended to those philosophers (particularly Robert Cummins) who have advocated a structural isomorphism TCD, or something close to it.

[6] At an earlier stage in the development of this work, I received a copy of Cummins (1996) in which he also adopts the approach just described. After I settled down a bit (it is nice to pioneer certain ideas, especially in the early stages of a career) I was pleased with the convergence of views. Indeed, even though the course of the current work was already established, my thinking on the problem of misrepresentation has benefited from Cummins' book.

The problem of misrepresentation will be my primary concern for much of the early part of this work.

The second more particular reason why I think that the resemblance theory deserves a thorough re-examination has to do with the emergence of *connectionism* (the computational theory of mind based on the parallel distributed processing, or just PDP, computational framework), in challenge to the hegemony of classicism (the computational theory of mind based on conventional, or digital, computational theory). I aim to demonstrate that while classicism is neutral in regard to the various TsCD available, connectionism, in contrast, suggests that a re-appraisal of the resemblance theory is in order. The reason is that PDP systems exhibit what is known as a "*semantic metric*". That is, physical similarity in the items doing the representing reflects similarity in the items being represented. This is strongly suggestive of the structural isomorphism TCD, something that has largely gone unrecognised. Connectionism, in other words, looks to have a natural affinity with structural isomorphism. This is an issue I intend to pursue in some detail in the latter part of the present work.

Overall, then, my aim is to motivate, develop, and defend the structural isomorphism TCD, as well as point out certain of its implications. The main topics are the doctrine of representationalism, the problem of misrepresentation, formulation and illustration of the structural isomorphism TCD, certain implications of the theory for our conception of mental content, and the relationship between content determination and cognitive science. These aims and topics, of course, overlap considerably.

Chapter One introduces both representationalism about the mind and the resemblance theory. First, I discuss the contemporary commitment to representationalism via an inroad through folk knowledge of the mind and cognitive science. I then seek to precisely locate the question of content determination within the representationalist framework. The question of content determination, I suggest, is to be located within a broader framework concerning the issue of representational status itself. Next, I make a first pass through the distinction between a first-order and a second-order resemblance theory. The context here is whether the resemblance theory is compatible with a naturalistic perspective. Chapter One closes with consideration of a foundational objection to the resemblance theory in general. Using material developed earlier in the chapter, I argue that this objection can be overcome. By the end of Chapter One, I hope to have put representationalism, the question of content determination, and the basis of the resemblance theory, on firm footing.

In Chapter Two I set about examining what I call the "orthodox approach" to content determination and misrepresentation. Such an approach is exemplified in the influential work of Fodor, Dretske, and Millikan. I will examine the differing TsCD these philosophers develop, as well as their various proposals for accommodating misrepresentation. I then detail the most telling objections encountered by these proposals, thereby revealing why it is widely thought that the problem of misrepresentation has yet to receive a satisfactory solution. The framework of Chapter Two is the character of the orthodox approach. I show that although there are certainly differences between the theories of Fodor, Dretske, and Millikan, they adopt essentially the same *type of approach* to content determination and misrepresentation. In my view, it is the nature of this strategy which has hampered the development of a natural means of accommodating misrepresentation. The orthodox approach will be subsequently contrasted with an alternative framework for content determination and misrepresentation.

The orthodox approach is pursued further in Chapter Three. Here the focus shifts (at least initially) from considerations of strategy, or methodology, to the conceptions of intentionality, content, and content determination which typify, perhaps even drive, the orthodox approach. I will put forward an alternative way of conceiving of such phenomena, as well as clarifying two kinds of representational error, namely, misrepresentation and selectivity. I attempt to isolate some factors responsible for the predominance of the orthodox approach. The overall aim is to lay-out the prevailing approach in order to reveal the character of a possible alternative strategy. The summary which closes Chapter Three highlights the fundamental differences between the orthodox approach and the alternative framework now in place.

The development and defence of the structural isomorphism TCD is the focus of Chapter Four. The structural isomorphism TCD I formulate emerges from consideration of various versions of the resemblance theory, including other possible interpretations of both second-order isomorphism in general, and structural isomorphism in particular. I develop a certain conception of mental content which naturally attaches itself to the structural isomorphism theory, a fine-grained conception I call "microcontent". I then formulate what is likely to be thought of as one of the most significant objections to the structural isomorphism TCD, the claim that it cannot secure determinate content. I show that there are various avenues of response to the "indeterminacy objection". Chapter Four closes with a discussion of how structural isomorphism operates within an approach to content determination and misrepresentation which differs significantly from the orthodox approach, one which naturally yields the capacity for misrepresentation.

Chapter Five has a number of overlapping concerns. The overall intention is to illustrate and defend the structural isomorphism TCD by applying it to certain more or less empirically based results coming from two domains of human representational activity, namely, temperature and colour. Within this broad aim resides a number of more specific, related concerns. First, I seek to show how the structural isomorphism TCD sits comfortably with, indeed explains, the *structured* character of our representational experience.[7] Our representational experience exhibits a certain structure which is to be explained by identifying representational experience in a given domain with a dedicated, physically-structured system of representation bearers. The physical structure exhibited by the type of representing system posited by structural isomorphism, that is, naturally accounts for the relationships obtaining among a system of representational experiences. The more contentious, and indeed surprising and perplexing view I want to defend is that the empirically flavoured results suggest that the contents of our mental states do not map onto *straightforward* features of the external world. Rather, the world of mental content is a *notional world.* I propose to sketch this idea rather than fully articulating it. I will (at least in this work) settle for pointing out that mental content is even more puzzling than we had previously supposed. What the findings more clearly illustrate, I think, is that the usual externalist view of content determination cannot be right. Rather, we must move to a thorough-going internalist, or narrow, account of content determination, one which structural isomorphism provides. In line with such an internalist account, I close Chapter Five by offering what I call a "projectivist" conception of intentionality.

---

[7] I have said representational "experience" here because the structured character of mental representation is most salient in the realm of phenomenal consciousness. As will become apparent, however, I think that mental representation in general has a structured nature.

Chapter Six enters the territory of cognitive science, the field founded on the assumption that mental processes are computational operations over neurally instantiated mental representation bearers. The main task of Chapter Six is a detailed investigation of both the classical and connectionist accounts of cognition, with particular focus on their respective relationships to the issue of content determination. The line of argument commences with an examination into the nature of computationalism in general, the result of which is a formulation of *generic* computationalism. Central to this formulation is the idea of a computational operation being, as I'll say, a *disciplined* operation. I will then argue that a crucial difference between classicism and connectionism concerns the differing manner in which the causal operation of conventional and PDP devices are disciplined. I aim to demonstrate that while the computational operation of conventional computers is not bound-up with the issue of content determination, PDP systems effect computational operations by employing structural isomorphisms between their representational substrates and task domains. Connectionists, that is, are committed to the structural isomorphism TCD. Connectionism emerges as a theory of cognition as a whole – that is, both cognitive processes and the contents of representation bearers whose transformations constitute cognitive processes – whereby central importance is accorded to the intrinsic physical properties of mental representation bearers.

In the brief Conclusion I rehearse the central theme of this work. Over the range of issues considered, the structural isomorphism TCD articulates a perspective on content determination that, while singularly unpopular in contemporary philosophy of mind, is nonetheless distinctive and viable: it is the physical properties of mental representation bearers themselves by which their content is conferred. In other words, by the close of the present work I hope to have shown that, contrary to popular opinion, the resemblance theory – in the form of structural isomorphism – is a coherent alternative to the orthodox theories, one which deserves to be seen as a serious contender in the quest to provide a naturalistic explanation of mental content determination.

CHAPTER ONE

REPRESENTATIONALISM AND RESEMBLANCE

Philosophical interest in intentionality goes back as least as far as Aristotle. The phenomenon was an important concern of Scholastic thought in the Middle Ages (indeed, the term 'intentionality' originated with the Scholastics). And though much thinking about the mind throughout the seventeenth and eighteenth centuries had an epistemological focus, reference to the property of intentionality is to be found, in one way or another, in the writings of Descartes, Locke, Hume, and Kant.[8] The term 'intentionality' was revived in the nineteenth century by Franz Brentano, who did much to legitimize the study of intentionality in its own right. Under the reign of behaviourism in this century, the idea that mental phenomena are "intrinsically" intentional fell into disrepute. Mental states, it was thought during this period, are only secondarily or derivatively intentional, with intentionality being primarily a property of the language used to characterise mental phenomena. However, such a view, the "linguistic analysis" of intentionality, is itself now very much out of favour.[9] These days the prevailing attitude (with one prominent exception to be touched on subsequently) is that mental phenomena are intentional in their own right. In the first section of this chapter, I'll explain why intentionality is taken to be vital to a contemporary scientific understanding of the mind.

As remarked in the Introduction, the contemporary practice is to view the naturalisation of mental representation in terms of answering the question of content determination. Indeed, one gets the impression from the literature that such a question *exhausts* the problem of naturalising mental representation. In section 2., I am going to suggest that the question of content determination is to be located within a broader issue concerning representational status itself. My target in this work is the question of content determination, but I will say a few words on how

[8] See Cummins (1989, Chp. 1) for a sketch of certain historical views on intentionality.

[9] There are a number of difficulties with such a linguistic analysis of intentionality (for an introduction, see Crane 1995, 31-37). The most damaging is that it is non-naturalistic, aiming to account for the intentionality of the mental in terms of a phenomenon (natural language) that is itself intentional. And according to many philosophers, the only viable account of the intentionality exhibited by natural languages (and other types of non-mental representational systems such as pictures, models, maps, and the like) is that it is *derived* from the intentionality of our mental states. It is minds which confer intentionality upon such cultural artefacts.

the issue of representational status might be addressed. I also propose to have a brief look at the distinction between original and derived intentionality.

In section 3., I sketch the difference between a first-order and a second-order resemblance theory, and point out why only a resemblance theory which operates with a second-order isomorphism is a viable option for explaining mental content. I then examine an influential objection to the resemblance theory, one which claims that it is fundamentally flawed due to the symmetric nature of resemblance. Though directed at the traditional first-order version of the theory, the objection also applies to a second-order rendering. I will argue that the objection rests on a confusion.

1. Why Representationalism?

The conviction predominating within the contemporary philosophy of mind community that the mind must be understood as a representational device derives, I think, from two quarters. One concerns philosophical reflection on folk knowledge of the mind. The other is the rise of cognitive science. I'll examine these in turn.

1.1. The Folk and Conscious Experience

One aspect of folk knowledge of the mind concerns our conscious experience. In undergoing certain types of conscious experiences – various types of perceptual experiences, in particular – we are presented with, made aware of, an external world; a world beyond, and independent of, ourselves. We are conscious of various features, objects, and states of affairs residing in an external reality. That is to say, certain conscious experiences are clearly representational states. This much is introspectively obvious to all of us. Indeed, it seems plausible to claim that all conscious experiences are representational. To undergo a conscious experience is to be made conscious *of something*.

Of course, the folk don't quite put it in such a way. The folk don't reflect on the nature of conscious experience and thereby arrive at the thesis that the mind is a representational system. But from at least the time of the Ancient Greeks, it has been apparent to many that have reflected on our everyday psychological knowledge that conscious experience is representational. And recent attempts to construct theories of consciousness have recognised this representational character. I should point out, however, that the idea that *all* experiences are representational is not universally accepted. Some philosophers hold that certain bodily experiences, such as of pain, are not representational. And Searle (1983, 1-2) argues that certain emotional states, such as feelings of undirected anxiety, do not exhibit intentionality. However, I think such views are mistaken. On the latter, it's possible to argue that Searle has mistaken the *diffuse* character of what is represented for a lack of intentionality. Emotional states, I think, are actually rather complex states, composed of a number of simpler representational elements. And in my view, experiences of pain point to certain disturbances either on the skin surface or in internal bodily parts. Here I think focus on the qualitative character of pains tends to obscure their representational component.[10] Along with a growing number of contemporary philosophers, then, I hold that while all representational states may not be conscious, all conscious states are representational.

Indeed, certain contemporary theorists have taken the representational perspective on consciousness one step further, developing so-called "representationalist" theories of consciousness (see, for example, Dretske 1993, 1995; Tye 1992, 1994, 1995). A representationalist theory of consciousness claims that the qualitative character of experience – the "raw feel" or "what-it-is-like" (Nagel 1974) – is to be exhaustively captured in representational terms.[11] Dretske has some nice statements concerning such an approach to consciousness: "Experiences…are conscious, not because you are conscious of them, but because, so to speak, you are conscious *with* them" (1993, 280-281). In being conscious, one is conscious of certain features and objects:

> An experience of x is conscious, not because one is aware of the experience, or aware that one is having it, but because, being a certain sort of representation, it makes one aware of the properties (of x) and objects (x itself) of which it is a representation (1993, 280).

[10] For a defence of the representational understanding of pain, see Tye (1995).

[11] Of course, one can recognise the representational character of experience without thereby being committed to a "representationalist" theory of consciousness. That is, one could assert that experiences have qualitative character over and above their representational aspects.

To undergo a conscious state, one is not (or, at least, not necessarily) conscious of one's mental state, rather one is conscious *of something*, that something the conscious state represents.[12]

1.2. Folk Psychology

Another aspect of folk knowledge about the mind is that doctrine philosophers refer to as "folk psychology" (alternatively, "commonsense psychology", "intentional psychology", or "belief/desire psychology"). Folk psychology is the cognitive resource we all employ in our everyday interactions with one another. The idea is that like the possession of commonsense knowledge regarding the behaviour of the physical world ("naïve physics", or "folk physics"), we have a commonsense body of knowledge concerning the behaviour of cognitive agents. This folk psychological knowledge is a predictive and explanatory framework which allows us to understand the behaviour of ourselves and others. Folk psychology is so deeply embedded in our conception of ourselves as cognitive agents, and so successful at facilitating our interactions with one another, that we are scarcely aware of its existence and operation. As Fodor (1987, 3) remarks, folk psychology "works so well it disappears." But all of us are capable of consciously applying folk psychology when required. For example, asked to explain the arrival of a group of students at a certain lecture theatre, I'd provide the following sort of folk psychological account. I'd ascribe to each of the students the desire for a reasonably prosperous and fulfilling working life, the belief that a university education enhances the prospects of achieving such a life, the belief that attending university lectures facilitates the passing of university courses, and the belief that the course in which they are currently enrolled is being conducted in the lecture theatre in which they are currently located.

One issue which interests philosophers is how folk psychology works. On this, there is something of a consensus. The operation of folk psychology rests with two principles: (1) the attribution of *information-bearing*, or contentful, *mental states*. (Typically the focus is on beliefs and desires, but folk psychological

[12] It might be objected that certain states, hallucinations for example, are conscious without there existing something one is conscious of. But surely one is conscious of something; the content of one's conscious state. It's just that what one is conscious of does not exist in the objective world as it currently stands. Such states are, of course, misrepresentations.

explanations also rely on the attribution of certain intentions, fears, memories, perceptual experiences, and so forth); (2) the attribution of certain kinds of *relations* between such mental states and the environment, other mental states, and behaviour. (Perceptual states, for example, are standardly caused by the external environment. Beliefs and desires enter into causal relations with one another to yield, for example, intentions. And intentions have a causal influence on the production of behaviour).[13] The connection between principle (1) and our concern with representationalism is obvious. Folk psychology is, for one thing, committed to the idea that mental states are *representational* states. As with the discussion of consciousness above, the folk don't quite put it this way. But it is clear that folk psychology embodies an understanding of the mind as a representation-using system.

However, the recognition that folk psychology operates with the positing of internal representational states is itself not sufficient to provoke a robust (philosophical) commitment to representationalism. There is a further issue, one which dominates philosophical discussions of folk psychology. Are the folk *correct* in their attribution of certain kinds of internal states with particular contents? *Why* does folk psychology work? Or, more broadly, what is the status of folk psychology?

According to many philosophers, folk psychology works because it rather accurately locks onto the nature of the mind's operation, and the nature of the relationship between mind and behaviour. Such a view constitutes a *realist* interpretation of folk psychology. Realists point to the enormous predictive and explanatory power of folk psychology. This success is itself to be explained, on their view, by interpreting folk psychology as a rudimentary *causal theory* of human behaviour. That is, the mental states folk psychology posits are to be understood as real, causally efficacious, internal representation bearers, and the transactions among these mental representation bearers, the environment, and behaviour are to be accepted as genuinely causal transactions. The commitment to a

[13] Often a third principle is cited. Folk psychology embodies a set of generalisations or *ceteris paribus* "laws", such as: "Cognitive subjects typically seek to enact their stated intentions."; and, "If a cognitive agent desires that *p*, and this desire is not over-ridden by other desires, and she believes that doing *K* will result in *p*, and she believes that behaviour *K* is in her power, then (*ceteris paribus*), she will perform *K*." While undoubtedly important to the everyday operation of folk psychology, such a set of generalisations is not a (separate) deep ontological principle of folk psychology. In effect, I think it's subsumed by principle (2).

realist interpretation of folk psychology is one factor which drives many contemporary philosophers to a representationalist conception of mind.[14]

Realism, it seems, has come to be the prevailing interpretation, with even the most prominent critics of folk psychology (such as Daniel Dennett and Paul Churchland) now moving away from their previously held antithetical positions.[15] Consequently, realism has emerged as a position which itself admits of various possible interpretations. At one end of the spectrum we have Jerry Fodor's "industrial strength" realism, where representation bearers are symbolic items written in the "language of thought". At the other end we can locate the weak realism defended in Daniel Dennett's more recent work (1987, 1991a). On this view, the mental states the folk posit are *abstracta* that serve as "indirect "measurements" of a reality diffused in the behavioural dispositions of the brain (and body)" (1991, 45). I suspect that many realists locate themselves, as I do, somewhere between Fodor and Dennett ("household grade" realism seems fitting).

---

[14] One of the most prominent and powerful defences of such a realist interpretation of folk psychology is to be found in Fodor (1987, Chp. 1). See also Horgan and Woodward (1985).

[15] The *instrumentalism* to which Dennett (1981, 1987) previously subscribed is the view that folk psychology is *not* a *causal theory* of human behaviour. On this interpretation, folk psychology does not lock onto the true nature of our cognitive operation, rather it is merely an instrument we employ to predict and explain behaviour. The mental states folk psychology posits are useful fictions visible from the "intentional stance" (Dennett 1987). The intentional stance is the practice of understanding an organism as an intentional system, one whose behaviour can be successfully predicted via the attribution of certain kinds of representational states. If a system's behaviour can be reliably predicted by utilising a certain such predictive scheme, then that is "*all there is*" to possessing certain representational states (Dennett 1987, 29). The representational states so ascribed are thereby "stance-dependent features" of the system (Baker 1989, 305). It is the stance-dependent understanding of folk-psychological states which marks the interpretation as instrumentalist. The shift in Dennett's thinking away from the instrumentalism of his earlier work to the weak realism he now defends has, I think, been largely driven by the difficulty of reconciling folk psychology's enormous predictive and explanatory power (something even Dennett does not dispute) with the claim that folk psychology is merely an instrumentalistic calculus, rather than an apparatus which rather accurately captures an underlying cognitive reality. See Fodor (1985/1990a, 6-7) for such an objection to instrumentalism. I'll have more to say on Dennett's weak realism in subsection 2.3. below.

For some time Paul Churchland was an advocate of *eliminative materialism* (see, for example, his 1979, 1981, and 1988). Unlike instrumentalists, eliminativists believe that folk psychology is a causal theory. In contrast to realists, however, they don't believe that folk psychology provides the foundations for a more systematic study of mind, because, in their view, it is a radically false conception of the causes of our behaviour. It's fair to say, however, that few philosophers have found Churchland's analysis of folk psychology convincing. Churchland's critics rightly point out that he mischaracterises folk psychology and overstates its defects (see, for example, Fodor 1987, Chp. 1; Horgan and Woodward 1985). Moreover, in his more recent work (especially his 1995), Churchland is up to his neck in *representational* talk, offering proposals for how the brain realises its representational capacities. Churchland has recognised, I think, that he cannot consistently maintain an eliminativist interpretation of folk psychology while addressing issues brought by a representationalist understanding of mind. The reason, of course, is that a fundamental principle of folk psychology, as pointed out above, is the attribution of *representational* states to cognitive agents. Churchland now refers to his position as *revisionary materialism*.

1.3. Cognitive Science

Like folk psychology construed in a realist fashion, cognitive science is targeted at the causal origins of our behaviour. But whereas folk psychology is a rather rudimentary theory in the service of our everyday social interaction, cognitive science explicitly aims for a precise account of the causal processes and mechanisms which underwrite our capacity to behave in an intelligent fashion in general. And the discipline of cognitive science is founded on the assumption that the causal processes mediating between environmental stimuli and behavioural responses are to be understood as *computational* processes. That is, mental processes are computational operations. Computational operations involve the processing of information, and the processing of information requires, of course, a means of realising information, a system of items which are the bearers of information or representational content. As Fodor has remarked, "computation presupposes a *medium* of computation: a *representational* system" (1975, 27, emphasis added). Computation just is the systematic manipulation of content-bearing items.[16] According to cognitive science, mental processes consist in the systematic manipulation of neurally-realised mental representation bearers. A commitment to computationalism, then, is a commitment to representationalism.

Why, however, accept the view that mental processes are computational processes? Is computationalism mandatory? As things currently stand, computationalism is for all intents and purposes indispensable since there simply exists no viable alternative. The idea that cognitive operations are computational operations is the only workable theory we have for explaining the causal processes underlying our intelligent behaviour.[17] Fodor often comments to the effect that viewing mental processes as computational processes is the best idea anyone has

[16] The notion of computation will be examined at greater length in Chapter Six.

[17] In recent years, some theorists have suggested that an explanation of intelligent behaviour can do without computation/representation (see, for example, Beer 1990, Brooks 1991, and Van Gelder 1995). In my view, however, Bechtel (1998), Clark (1998), and Clark and Toribio (1994) have decisively shown such a claim to be fundamentally flawed. The so-called "anti-representationalists", on closer examination, are really only arguing against the "symbolic" representationalism of the *classical* computational theory of mind. Indeed, a central aim of this thesis is to show that there is an alternative to the traditional symbolic conception of representation, an alternative that emerges once we begin to contemplate a resemblance based account of content determination.

had about the mind (see, for example, his 1992).[18] Cognitive science offers an answer to "a great mystery about the mind": How could a purely physical system (that is, the brain) exhibit the *rationality* required to produce the sort of intelligent behaviour exhibited by organisms such as ourselves? (Fodor 1987, 20). An intelligent system is one which is able to respond in a flexible manner to changing environmental circumstances such that its behaviour is appropriate to its conditions. And it looks for all the world that what this requires is, at bottom, that the system possess information concerning the state of the environment (including its past experience in its environment), and that such information shapes the system's behavioural responses. Another way of putting this, of course, is to say that the system must possess internal *representational* items which enter into causal transactions with other internal states and behaviour in a *content-relevant* fashion. Or, to express it in terminology which will be important to the deeper excursion into cognitive science conducted in Chapter Six, an intelligent system's behaviour must be driven by causal processes which are "semantically coherent".[19] A system whose causal operation is somehow disciplined such that its behaviour-causing items have content-relevant effects is a rational physical system. Cognitive science, then, shows how the *mechanisation* of rationality is possible, since computational operations are disciplined physical operations involving the semantically coherent or content-relevant manipulation and transformation of representation bearers.

One thing which goes unnoticed too often is the striking convergence between computationalism (construed generically, as I've been doing) and a realist interpretation of folk psychology.[20] Principle (2) of folk psychology – the ascription of certain kinds of causal relations into which mental representation bearers enter – is of particular importance here. On a realist interpretation of folk psychology, recall, mental states enter into genuine causal relations in just the way folk psychology says, namely, in a manner concordant with their representational contents. Perceptual states, for example, are typically caused (at least in part) by the environmental conditions such states are about. Intentions and desires, on the other

---

[18] Of course, Fodor's remarks pertain to Turing and the idea that mental processes are *digital* computational processes. I think, however, that we can recast Fodor's comment in generic terms – in, that is, a way neutral between digital and PDP-style computation – without it losing its force. (Fodor, though, would be *mortified*).

[19] I'll discuss the idea of semantic coherence at greater length in Chapter Six.

[20] But see Fodor (1987, Chp. 1) where the convergence is rendered explicit (although, of course, Fodor makes the point in regard to classicism, rather than generic computationalism).

That completes the examination into the reasons for the contemporary commitment to mental representation. I have one subsidiary, and very brief, point to make by way of closing this section. It concerns the kind of representational capabilities at which theories of content determination are aimed.

1.4. The Representational Base

Given that it's early days for the naturalisation of intentionality, proposals don't set about trying to explain all of our representational capabilities in one fell swoop. Rather, they are concerned to pay-off the principal, and worry about the credit charges later. That is, current theories are clearly pitched at our more primitive representational faculties, what we might call the "representational base". This is the kind of representational activity we share with other, simpler animals, the sort of activity involved in representing one's immediate environment.

The reason for such a starting-point is obvious, but what is not so apparent is that the representational capability targeted operates as a constraint on any proposed theories. We want a way of conferring content Nature would have selected for all those creatures with similar sensory systems, and one which will operate across various sensory modalities. More stringently, any proposal cannot rely (even in part) on other mental capabilities only we possess (such as the capacity for natural language). The hope is that a theory which successfully explains our perceptual and sensory representational abilities will eventually be shown to scale-up to account for those more complex capacities Nature presumably erected on (or around) the representational foundations. Or perhaps a successful account of the representational base will require augmentation with a different theory to explain our more sophisticated representational faculties.

In any event, a theory which dealt successfully with the representational base would break the back of the problem of mental content; it would solve the *mystery* of mental content. The theory to be subsequently developed in this work is explicitly targeted at explaining our more basic modes of mental representation.

Now that I've set out the motivation behind representationalism, and its general aspirations, it's time to examine more closely just what it is about mental representation that requires explanation.

2. Representationalism: Two Questions, Two Theories

Representationalism is the general doctrine concerned with a number of issues arising from its assumption that the mind is a representational device. In the Introduction, I pointed out that 'representation' is ambiguous between the item doing the representing and the item being represented. Consequently, I introduced the distinction between, respectively, the "representation bearer" and the "representational object". In this section, I'm going to suggest that 'representation' is ambiguous in a further way, and so introduce another distinction. The distinction yields what I take to be the two central issues of representationalism.

2.1. Two Questions, Two Theories

If representationalism is to be a successful doctrine it is required to provide an account of how mental representation is possible. But just what is it that is in need of explanation? A little reflection suggests that there are *two* central things concerning mental representation which require explanation. First, we require an account of how the states of, or items in, a physical system (the brain) could *represent* – that is, exhibit intentionality – to begin with. The question arises because, of course, not everything has representational states. Originally intentional states are something that only relatively few types of entities possess. Representational systems are thought to be a special class of system. Only those systems capable of *intelligent* behaviour are deemed to possess internal states which are intentional in their own right (see Fodor 1986).

Second, a mental representation bearer represents something in particular, be it a state of affairs, an object or objects, a class of things, a feature or property of an object, or a value of a variable. (For the most part, in what follows I'll simply employ the term "representational *object*" to refer to that which is being represented). Representation bearers, that is, possess specific aboutness, or representational *content*. We require an account of how representational content is conferred, or determined.

There are, then, two crucial questions we might ask about mental representation:

(1) By virtue of what do mental states *represent*, or possess intentionality?

(2) By virtue of what do mental states possess representational *content*?

We might call the first question the "question of representation", and a theory which addresses this question a "theory of representation". A theory of representation says what constitutes the property of "aboutness"; it states the conditions required to generate "aboutness". Such a theory, in effect, tells us what makes an internal entity a mental *representation bearer*.

Employing terminology already introduced above, we might call our second question the "question of content determination" (QCD), and that component of representationalism which addresses this question a "theory of content determination" (TCD). A TCD, of course, purports to explain how content is conferred upon mental representation bearers.[22]

The naturalisation of intentionality, I submit, requires answers to both the question of representation and the question of content determination.[23] As suggested, however, the project is not currently viewed in such a way. Overwhelmingly, writings on mental representation focus on what I've called the question of content determination, to the relative neglect of the question of representation. Indeed, current terminology would suggest that the prevailing attitude is that the QCD *exhausts* the problem of naturalising intentionality. What are, in fact, theories of content determination are often referred to as theories of "mental representation". And across the field, one finds the terms a "theory of mental representation" and a "theory of mental content" employed synonymously. This terminological practice is, at best, misleading. More seriously, it often indicates a conflation of the issues of representation and content determination. It suggests that all there is to the naturalisation of intentionality is a satisfactory TCD. Such a view gives rise to certain serious misconceptions which I will describe, and attempt to clear up, a little further on (see subsec. 3.2.).

Though the question of representation and the question of content determination are distinct issues, and a theory of representation and a theory of

---

[22] An analysis of the notion of representation along the lines just rehearsed is to be found in Von Eckardt (1993, Chps. 4-8). Indeed, I am indebted to Von Eckard for clarifying my thinking on what is required of an explanation of representation. See also Cummins (1989, 10; 1996, 124).

[23] In my view, satisfactory answers to both questions would provide for a vindication of principle (1) of folk psychology (see subsec. 1.2.).

however, I propose to have a look at the question of representation, and, in particular, Peirce's notion of the interpretant.

2.2. The Question of Representation: A Peircean Proposal

It is the inclusion of the interpretant which distinguishes Peirce's account from the typical conception of representation as merely a dyadic relation obtaining between a representation bearer and its representational object. On Peirce's theory, something is a representation bearer if it *operates*, or functions, *as* a representation bearer *for* an interpreter. An item operates as a representation bearer for a cognitive agent by producing an interpretant in the agent. For Peirce, the interpretant is the mental effect the representation bearer has on the interpreter, such that the interpreter is brought into relation with the representational object. In the case of *non-mental* representation bearers, such a mental effect is a thought, or sequence of thoughts, in the mind of the interpreter. So, in regard to the representation bearer-interpretant relation, a representation bearer must have the power, or capacity, to produce a mental effect in an interpreter (1.542). And in regard to the interpretant-representational object relation, because the interpretant brings the cognitive subject into relation with the representational object, the cognitive subject must in some sense be capable of understanding the representation bearer (2.228, 4.536, 5.287). In common with many contemporary philosophers of mind, Peirce held that thoughts are representational states (5.283, 6.338). A thought – the interpretant of a non-mental representation bearer – brings a cognitive subject into relation with the bearer's representational object because the thought is itself *about* the representational object.

Patently, however, characterising the interpretant as a thought won't do for mental representation. It produces an infinite regress. The interpretant produced by a mental representation bearer, if understood as a thought, is itself a mental representation bearer, and therefore in need of interpretation, which in turn implicates another mental representation bearer, and so on.[25] Peirce himself was aware of this difficulty, and sought to unpack further the notion of the interpretant

[25] Indeed, it can't be the whole story for non-mental representation either. Because the interpretant of a non-mental representation bearer is a mental representation bearer, itself requiring interpretation, then if we do not characterise the interpretant in a way which does not implicate further representation bearers, non-mental representation bearers will also lack interpretation.

in terms of the causal powers the mental representation bearer possesses (5.476; see also Von Eckardt 1993, 283-284). That is, the mental representation bearer enters into causal transactions with other mental representation bearers, and ultimately, the cognitive subject's behaviour. At bottom, then, the process of interpretation consists in a *modification of the subject's behavioural dispositions*. The mere *presence* of mental representation bearers *qua* neurophysiological structures modifies a subject's dispositions towards behaviour. The idea of modifications to a subject's behavioural dispositions yields an account of the interpretant which does not itself involve intentional notions. And again, because the interpretation process brings the cognitive subject into relation with the representational object, the modification of the subject's behavioural dispositions must be dependent on, or in accordance with, the representational object concerned. In this way, the process of interpretation is one whereby the mental representation bearer so modifies the subject's behavioural dispositions that she is disposed to behave in a manner which is (in some way) appropriate to the representational object. A given mental representation bearer *qua* neurophysiological structure might effect its particular modification of behavioural dispositions in accordance with, for example, its "shape". Along such lines, I provide a connectionist account of the content-appropriate causal powers of mental representation bearers in Chapter Six.

So, a Peirce-type answer to the question of representation suggests that an internal state of a cognitive creature is a mental representation bearer – exhibits intentionality – if it bears some kind of relation to a representational object, and effects a modification of the subject's behavioural dispositions toward the representational object. I suggested above that a distinction between what I've called the question of representation and the question of content determination is seldom explicitly drawn, and that contemporary writings on mental representation overwhelmingly focus on the latter question. That noted, however, I do think many contemporary philosophers would agree with a Peircean answer to the question of representation. What little that has been said on the issue of representational status suggests that, for certain philosophers, it is something like the *role* of a mental item in the overall cognitive economy that makes it a mental representation bearer.

Most explicitly perhaps, Dennett writes: "What makes some internal feature of a thing a representation could only be its role in regulating the behaviour of an intentional system" (1987, 32, emphasis deleted). Millikan (1984, 1993) has long argued for the importance of devices which purportedly "consume" or use representations in addition to those mechanisms which produce representations: "...what makes a thing into an inner representation is, near enough, that its function is to represent. But...the way to unpack this insight is to focus on representation *consumption*, rather than representation production. It is the devices that *use* representations that determine these to be representations...."(1993, 85-86). Clark suggests: "The status of an inner state as a representation...depends not so much on its detailed nature (e.g., whether it is like a word in an inner language, or an image, or something else entirely) as on the role it plays within the system" (1997, 146). And Lloyd writes: "...the requirement that representations have a cognitive role in the systems which contain them is also a capacity that may go unrealized in specific cases. Particular representations may evaporate without a trace. What *is* necessary, however, is that the system in which they occurred was one in which they *could* have made a difference, either to other representations or to behavior" (1989, 18, emphasis added; see also 63-67). Such statements suggest that the effect (or potential effect) of an internal item is (at least in part) what makes for mental representation, an idea which, as outlined, is present in Peirce's analysis of representation.

2.3. The Question of Content Determination: Original and Derived Intentionality

As with Peirce's analysis, in contemporary discussions the QCD is typically construed as one concerned with the nature of the relation obtaining between mental representation bearers and objects in the represented world. Theories of content determination differ in terms of the account they give of the relation purported to hold between representation bearers and their representational objects. Causal theorists say that the nature of the relation is, of course, causal. Resemblance theorists claim that the relation is one of (some kind of) isomorphism. Or, to be precise, TsCD which are consistent with the doctrine of *original intentionality* differ in regard to the proposed nature of the relation between mental states and representational objects. One prominent TCD – Dennett's (1987) neobehaviourist account – does not fall within this doctrine Dennett has famously argued that mental states exhibit only *derived*, not original, intentionality.

The doctrine of original intentionality is the view that mental states are originally and determinately contentful. To say that mental states are *originally* contentful means that they have content objectively, or in and of themselves. Their content is not conferred as a result of the interpretive (predictive and explanatory) interests of other cognitive systems. They are contentful in their own right. To hold that mental states are *determinately* contentful is to assert that there is a fact of the matter as to the exact (and discrete) content of any given mental representation bearer. That is, any mental representation bearer genuinely has a definite content (even if, in practice – by, for example, observing a cognitive agent's behaviour, or asking her to introspect – we cannot fix upon the exact content).

In claiming that mental states exhibit only *derived* intentionality, Dennett disputes the ideas of original and determinate content. For Dennett, internal items possess content only insofar as a subject's behaviour can be successfully interpreted via the adoption of the intentional stance and the attribution of certain contentful states to that subject. Mental states are not contentful in their own right; their content is a product of a certain interpretive scheme. Unlike his instrumentalism, however, where content is an entirely stance-dependent feature of organisms, Dennett's weak realism emphasises that intentional ascription tracks *objective* patterns in behaviour. Yet while such patterns are really there, they are only detectable from the perspective of the intentional stance, a vantage point we naturally take-up to generate interpretations of others' behaviour. Attributions of content to internal states can only be legitimately made on the basis of the patterns of behaviour an organism exhibits, and such attributions are made from the vantage point of the intentional stance which we adopt to serve our interpretive needs (see Dennett 1987, 39). And this means we can only ever legitimately talk of the contents of internal states in the sense that it is *as if* mental states themselves exhibit such contents. To again quote a remark which goes to the heart of Dennett's position, intentional ascriptions serve as "indirect "measurements" of a reality diffused in the behavioural dispositions of the brain (and body)" (1991a, 45). Mental content, in a sense, consists in, or is "realised" by, these complexes of behavioural dispositions. And so, for Dennett, the original intentionality of mental states is a myth.

A consequence of Dennett's account – and one he readily embraces – is that it may turn out that different interpreters, or different interpretational schemes employed by the one interpreter, ascribe *different* sets of contentful states to the one cognitive subject, and yet the two schemes happen to be *equally* predictively successful. In such cases, according to Dennett (1987, 28-29, 40-41), the content of the subject's representational states is indeterminate:

> ...I...maintain that when these objective patterns [in behaviour] fall short of perfection, as they always must, there will be uninterpretable gaps; it is always possible in principle for rival intentional stance interpretations of those patterns to tie for first place, so that no further fact could settle what the intentional system in question *really* believed (1987, 40).

And as he points out, a crucial difference between those who defend original intentionality (such as Fodor, Dretske, and Searle), and his own position, is the issue of whether mental content is ultimately determinate or indeterminate: "...Fodor and company...suppose that, independently of the power of any observer or interpreter to discover it, there is always a fact of the matter about what a person (or a person's mental state) *really means*" (1987, 294). According to Dennett, because content resides in those complexes of behavioural dispositions distributed across a cognitive subject's brain and body, there is no deeper fact concerning the "real" contents of a subject's internal representational states. To ask of the objective contents of a subject's internal states is, for Dennett, an empty question.

The differences between original intentionality and derived intentionality highlighted above, I want to suggest, can ultimately be traced to a fundamental point of disagreement. The dispute rests with the question of whether content determination is a matter of an *objective relation* between representation bearers and representational objects. According to those TsCD which hold that mental states possess *original* intentionality, the determination of mental content consists in the existence of an objective relation between mental representation bearers and the objects they represent. Debate within this doctrine concerns the nature of such an objective relation (is it causal, one of resemblance, or some other kind?). Proponents of *derived* intentionality, such as Dennett, deny the existence of any kind of objective relation holding between an organism's internal states and the environmental objects the organism represents. It is *this* difference, in my view, which marks the great divide between original and derived intentionality, and indeed, constitutes the source of many of the disagreements between Fodor (and the like) and Dennett (and his followers) as regards mental content. It is because some kind of objective relation is posited to hold between mental representation bearers and representational objects that mental states are claimed to be originally contentful. For the very same reason, original-intentionality TsCD hold that mental states have determinate content. Such objective content is independent of the interpretive interests of other cognitive systems. And it is precisely because of the rejection of an objective relation between internal states and environmental objects that Dennett, in defending derived intentionality, claims that content determination is solely a matter of a stance-dependent, content-interpretation of a subject's behavioural patterns. Since there is no objective relation which determines the content of internal states, there is no deeper fact as to the "real" content of such internal states when rival content-attributing schemes are equally successful in interpreting a subject's behaviour.

The resemblance TCD being developed in this work holds that mental states possess original intentionality, a claim I will briefly reinforce below. In the next section I take a preliminary look at the resemblance TCD, and in particular, demonstrate how two initial problems it confronts can be overcome.

3. Resemblance: Naturalism and Symmetry

As I pointed out in introducing this work, there is one thing upon which the overwhelming majority of contemporary philosophers of mind agree: the resemblance theory is not a fit candidate for the question of content determination. In this section I'll investigate two considerations which, in my view, are largely responsible for such an attitude. The first suggests that the resemblance theory is incompatible with the prevailing materialist conception of mind. The second claims that the resemblance theory is "logically" flawed. I'm going to argue that neither consideration provides a good reason for rejecting the resemblance theory.

I should emphasise that the present section is merely a *first pass* at the resemblance theory. I will supply only that detail required for my primary concern here, namely, showing that certain foundational considerations brought against the resemblance TCD are shortsighted. The resemblance theory, particularly those versions which operate with a second-order isomorphism, will be examined in far more detail in Chapter Four.

3.1. Resemblance: First-Order, Second-Order, and Naturalism

On the resemblance theory, content determination is a matter of (some sort of) resemblance (similarity, or likeness) between representation bearers and representational objects. That is, a representation bearer represents what it represents by virtue of bearing a likeness, in some way, to its representational object. The representation bearer (or system of representation bearers) reproduces some feature or features of the representational object (or system of representational objects). Representation bearer and representational object have something in common. (Obviously, everything has something in common with any other thing, but I'll sharpen the characterisation below).

What I've referred to as the "traditional" resemblance theory operates with a "*first-order*" (Shepard and Chipman 1970) resemblance relation or isomorphism. First-order resemblance is to be understood in terms of *physically equivalent* features. A representation bearer represents its representational object by virtue of having physical properties in common with the representational object. A first-order resemblance relation has long been thought to underwrite the conferral of non-mental pictorial content. The idea that a ("realist", or "representational") picture represents that which it resembles is intuitively appealing, even obvious.[26] A realist painting represents the blue of the sky by blue pigment; the spatial arrangement of the representational object, or scene, is mirrored by the spatial arrangement in the painting. A pictorial representation bearer and its representational object sustain a resemblance relation in terms of their *appearance* or "visually discernible" (Sartwell 1994) properties. Furthermore, the resemblance theory of pictorial content should be characterised as adverting to those visually discernible properties which belong to a picture's representational substrate. The "representational substrate" is to be distinguished from the "material substrate".[27] The material substrate consists of all the physical properties possessed by a representation bearer. The representational substrate is a subset of the material substrate. It consists of those properties which are representationally significant; those properties, that is, which are "interpreted" as being about some object (of

[26] However, even in the realm of non-mental pictorial content, the resemblance theory is not without its detractors. One of the more influential critiques is to be found in Nelson Goodman's (1969) *Languages of Art*.

[27] I owe these terms to Gerard O'Brien (personal communication).

which spatial arrangement, proportionality of parts, and colour properties would appear to be likely candidates in the domain of pictorial representation). The moral here generalises: any version of the resemblance theory, for any proposed mode of representation, must specify those dimensions or variables of the representing and represented environments between which the purported resemblance obtains.

However, a resemblance theory that operates with a first-order isomorphism can only be sustained as long as one cleaves to a dualist conception of mind. And (for very much the most part) contemporary philosophy of mind is committed to a thoroughly materialist perspective. The reason why a materialist must reject out of hand the first-order resemblance theory is, of course, that the brain simply does not possess a range of properties sufficient to reproduce all of the features which can be represented. Obviously, neurophysiological structures are not, for example, green, so the brain can not represent green by tokening a green representation bearer. Nor are neurophysiological structures shaped in an appropriate way, such that they could straightforwardly resemble the shapes we encounter in our environments.[28] And how is a neurophysiological structure supposed to resemble, in a first-order fashion, the aromatic properties of a 1951 Penfolds Grange? (The obvious difficulty of getting your neurophysiological structures in close proximity of a 1951 Penfolds Grange notwithstanding). The straightforward inconsistency of the first-order resemblance theory with materialism is one reason why many philosophers simply dismiss the idea that mental content determination is founded on resemblance.

However, it is not necessary that a resemblance theory operate with a first-order isomorphism. A "*second-order*" (Shepard and Chipman 1970) conception of the resemblance relation is possible. What is required to secure a second-order isomorphism is that variations in a feature of the represented domain are tracked by *corresponding* variations in a feature of the representing domain. Or, in slightly different terms, the various possible values of a represented variable are systematically mirrored by the values of a property possessed by the representing domain. On second-order isomorphism, the representing domain reflects the

---

[28] There is a rather crude first-order resemblance between activation in that area of the visual cortex known as V1 and spatial features of perceived environmental objects. Obviously, however, such a superficial resemblance will not make for a general theory of mental content determination.

represented domain's *relational structure*. Content is determined by virtue of a *structural complementarity* of representing and represented domains.[29]

I will discuss second-order isomorphism in much more detail subsequently (Chp. Four). For now, the concern is with how it allows the resemblance theory to overcome the materialism objection. A second-order resemblance theory is one which results from a weakening of the resemblance relation such that rather than representation bearers reproducing physically equivalent properties of representational objects, representation bearers employ properties which are physically *non-equivalent* to their representational objects. This frees the representing domain – neural systems – of having to possess that vast range of properties the world instantiates, that enormous diversity of features mental states are able to represent. Neural representation bearers need not have physical properties in common with their representational objects in order for a resemblance relation to hold between the brain and the representable world. A second-order resemblance relation can hold, one which obtains between the structure among a system of neural representation bearers and the structure among a system of representational objects. So the resemblance theory, at least in the form of a second-order isomorphism, is compatible with a materialist conception of mind. It might be thought, however, that the resemblance theory now faces a materialist version of the original problem. Is it plausible to suppose that the brain exhibits enough variability to mirror the structure of all the representational domains creatures such as ourselves encounter? Though not aiming the discussion directly at this question, I will in effect subsequently show that the brain's representation bearers (activation patterns across neural networks on the connectionist approach I adopt) do exhibit sufficient variability in terms of, roughly, their "shape", to cope

[29] The term 'structure' serves a dual role in this thesis. As I will demonstrate in Chapter Four below, there are in fact *two* kinds of second-order isomorphism, which I will call "functional" isomorphism and "structural" isomorphism. As versions of second-order isomorphism, both are broadly characterised in terms of a sharing of relational *"structure"*. (How they are to be differentially characterised is something I will spell out in Chp. Four, sec. 1. below). On the one hand, then, "structure" is used in a *general* fashion, one which subsumes both functional and structural isomorphism. On the other hand, "structure" is used in a more *specific* manner – as in "structural isomorphism" and its characterisations – which refers only to that version of second-order isomorphism utilising a structural isomorphism. How the term is being employed will typically be clear from the context. I'll write of "second-order isomorphism" when I mean to refer to *both* functional and structural isomorphism (so the word 'structure' here is being used in the *general* fashion), and "structural isomorphism" when I mean to refer to that particular version of second-order isomorphism (so 'structure' here is being employed in the more *specific* manner).

with the diversity of items in the representable world (see Chp. Six, and also Chp. Five, note 108).

There is another problem (roughly the converse of the one just touched on). If the structure of a given neural representing domain mirrors the structure of a represented domain, then surely it's also going to mirror the structure of a *whole host* of represented domains, meaning that the second-order resemblance theory results in radical content indeterminacy. This is indeed an important objection to the theory being proposed here, and I'll investigate various possible answers in Chapter Four, following the considerably more detailed investigation into a second-order isomorphism account of mental content determination.

Before moving on to critically examine the second of two preliminary reasons for rejecting the resemblance theory, there is a subsidiary issue I should briefly consider. It is sometimes objected that the resemblance theory is not compatible with a naturalistic framework in another way. Fodor (1984/1990a, 33), for example, worries about whether resemblance is in fact a natural, or objective, relation (one, that is, which can hold independently of the interpretive or recognitional capacities of cognitive agents), or whether it is merely, as he says, "in the eye of the beholder". Fodor contrasts resemblance with causation, the latter, in his view, a natural relation *par excellence.*[30] Surely, however, resemblance *is* an objective or natural relation. Two items can have properties in common (first-order isomorphism), or two variables can have value-structures in common (second-order isomorphism), without such resemblance relations being *conferred* by the recognitional capacities of cognitive agents.

What *is* true is that in the case of *non-mental* representation a cognitive agent must perceive or recognise the respects in which a representation bearer resembles its representational object, such that the cognitive agent can be brought into relation with the representational object. That is, the cognitive agent must in some way understand or interpret the representation bearer in order for the picture (for example) to *function*, or operate, *as* a representation bearer *for* the cognitive agent. Nothing in the picture serves a representational function for the picture. Pictures serve as representation bearers for us, and so the resemblance must be noticed. But this is something we knew all along: The operation of non-mental

[30] See also Cummins (1989, 31-32) for a discussion of the idea that similarity is "perceived similarity".

representation bearers is mediated by mental representation bearers (perceptual experiences, for example). Moreover, unless there does obtain some sharing of properties between the representation bearer and its representational object, the cognitive agent would not be brought into relation with the representational object, and the representation bearer would thereby not operate as a representation bearer. A drawing of a cat represents a cat to us because the texture of the ink marks produces an experience roughly similar to that of perceiving a real cat's fur. But we wouldn't even get this far were it not for certain objective similarities between the drawn cat's shape and the shape and arrangement of body parts of a real cat. Again, the resemblance between the drawing and the representational object must be noticed (or noticeable) in the case of non-mental representation in order for the drawing to be a representational item. Notice, however, that this sort of thing is *standard* for non-mental representation, regardless of the nature of the content-conferring relation. The *causal* relation between a doorbell ringing and the presence of someone at the door must be recognised in order for the "ding-dong" to operate as a representation bearer. Invite a Martian into your house (a Martian here is anyone not aware of the correlation between doorbells ringing and someone being at the door), have someone ring the doorbell, and observe. Won't the Martian just *sit there*? The worry that resemblance is not a natural relation is born out of focusing on the natural home of the first-order resemblance theory, namely, non-mental (pictorial) representation. And it's a product of conflating the determination of content with the recognition of content by cognitive agents.

Consider now mental representation and second-order resemblance. Suppose, for a moment, that a second-order resemblance TCD is true of mental representation, and it is differences in the "shape" of mental representation bearers *qua* neurophysiological structures that is the variable physical property by which the representing domain represents the represented domain. Then, suppose that the mere presence of a representation bearer (the mere "tokening") will effect a modification of behavioural dispositions (a modification in accordance with the representation bearer's shape). The "interpretation" here (part of that which makes the neurophysiological structure a mental representation bearer) is, as suggested above (sec. 2.), the modification of behavioural dispositions. The cognitive subject is disposed to behave toward the representational object because of the representation bearer's content. *No recognition* of the dimension of the representing domain by which a resemblance relation obtains is required of the cognitive agent harbouring the representing system.

So, having circumvented any worries concerning whether the resemblance theory is compatible with the naturalistic framework, it's time to turn to the second main reason for the current unpopularity of the resemblance TCD.

3.2. The Symmetry-Asymmetry Objection

It might be thought that *any* version of the resemblance theory, whether first-order or second-order, or applied to the non-mental or mental realm, is unworkable. A well-known objection due to Nelson Goodman (1969) strikes at the very foundation of the resemblance theory. As Goodman correctly observes, resemblance is a *symmetrical* relation, while representation is an *asymmetrical* relation. If *A* resembles *B*, then *B* also resembles *A*. But if *A* represents *B*, *B* does not (typically) represent *A*. Consider Goodman's own example. A painting represents the Duke of Wellington, the resemblance theory would have it, because the painting resembles the Duke of Wellington. The Duke, however, resembles the painting to the same degree as the painting resembles the Duke. But while the painting represents the Duke, the Duke does not represent the painting (Goodman 1969, 3-4). Resemblance is "logically" unsuitable as an account of representation because it does not exhibit the asymmetry required of representation. Though directed at the first-order resemblance theory, the objection applies, I think, to any theory founded on resemblance. An advocate of resemblance, in either form, must therefore confront the symmetry problem.

Goodman's symmetry-asymmetry objection, somewhat surprisingly given the intuitive appeal of resemblance in regard to (non-mental) pictorial content, has been subject to little critical attention. Instead, it is usually cited with unquestioning acceptance, and its influence has ranged wide. Jerry Fodor, for example, writing on *mental* representation, believes it shows the resemblance theory to be a non-starter: "...the resemblance theory...is certainly wrong....[R]esemblance is a symmetrical relation, and representation isn't; so resemblance can't *be* representation" (1984/1990a, 33). Goodman himself would appear to think that it is sufficient cause to reject the resemblance theory. However, as I've previously argued (Files 1996), the symmetry-asymmetry objection rests on a fundamental confusion. It involves a conflation of the question of representation and the question of content determination.[31]

Goodman correctly notes that resemblance exists *without* representation. For example, cars coming off an assembly line, he observes, resemble one another very closely, but none is a representation of any of the others (1969, 4). Two trees might resemble one another, but neither is a representation of the other. Goodman's symmetry-asymmetry objection, I've suggested (Files 1996, 404), is on a par with such cases: A subject resembles, but does not represent, her portrait. That is, the symmetry of resemblance is simply a special case of *resemblance without representation*. From the observation that there can be resemblance without representation it follows that "resemblance…is *no sufficient condition* for representation" (Goodman 1969, 4, my emphasis). And for Goodman, this shows that the resemblance theory should be rejected.

However, the move from the correct claim that resemblance is not sufficient for representation to a rejection of the resemblance theory is, as suggested, based on a conflation of the questions of representation and content determination. That is, it is due to a failure to recognise that, in addition to the question of content determination, there is another issue, that of representational status. The problem, as I think Goodman sees it, is that two items can resemble one another in non-representational, as well as representational, circumstances. And as there is nothing in *resemblance itself* to distinguish these two sorts of cases, we should reject the resemblance theory. Having conflated representation and content determination, it is inevitable that Goodman should arrive at such a view. The conflation is revealed in Goodman's characterisation of the resemblance theory as representation occurring "if and only if" resemblance obtains. And in the list of "faults", Goodman straightforwardly lines up resemblance with representation (1969, 4). Resemblance is treated as offering a complete account of how representation is secured; it is expected to provide the conditions required for representational status. An appreciation of the real reason why resemblance is insufficient for representation, however, shows that dismissing the resemblance theory is a mistake. The resemblance theory is an answer to the question of *content determination*, not the question of representation, and any TCD (as suggested in

[31] The symmetry-asymmetry objection is briefly critically examined by Blinder (1986), Gilman (1972), and Poole (1974). Their (to my mind, unsuccessful) responses differ from the one I develop.

sec. 2.) must be embedded within a theory of representation. A resemblance-based account of content determination cannot itself say what makes certain instances of resemblance content conferring. Goodman's criticism of the resemblance theory in this regard therefore constitutes an illegitimate demand. It is the theory of representation which distinguishes content-conferring resemblances from non-content-conferring ones by supplying the conditions required for representational status. And in so doing it is the *theory of representation* which provides for the required asymmetry of representation.

Of course, a complete reply to Goodman's objection requires an answer to the question of representation. But I've already gone some way on such an issue. In section 2., following Peirce, I suggested that a crucial component of a theory of representation is the notion of the interpretant. The painting of the Duke of Wellington produces (or has the power to produce) an interpretant which brings a cognitive agent into relation with the representational object, namely the Duke himself. The Duke, on the other hand, does not generate an interpretant, and so does not bring a cognitive agent into relation with the object to which he is related by resemblance, namely the painting. The painting represents by virtue of being taken by cognitive agents as being about something; the painting, but not the Duke, functions as a representation bearer for cognitive agents.[32] And in the realm of mental representation, an internal item is a representation bearer because it has the capacity to modify a subject's behavioural dispositions toward the representational object, the object to which the representation bearer is related by, for example, resemblance.

The plausibility of the response to Goodman's objection I've offered here, and, indeed, the Peircean perspective on representation I've favoured in general, is reinforced by the fact that it affords a straightforward response to a problem which besets the *causal* theory, namely, "pansemanticism" (see, for example, Fodor 1990a, 92-93)[33]. Now, I'm not ordinarily in the habit of seeking to serve the interests of the causal theory, but I do think there's an interesting parallel between

---

[32] I've previously argued (Files 1996, 406-409) that Goodman incorrectly characterises the role of *convention* in non-mental pictorial representation. Goodman advocates a conventionalist theory of pictorial *content determination*. I suggest, however, that convention is properly placed in the theory of representation. Pictures (and not Dukes, for example) function as representation bearers because it is a convention that pictures represent things (including people) while people do not represent pictures. Convention, located in the theory of representation, governs the interpretation of certain artefacts *qua representation bearers*.

[33] I discuss the causal theory in more detail in Chapter Two.

the problem of pansemanticism and the symmetry-asymmetry objection. The problem of pansemanticism is that the causal theory threatens to make representational items out of those we believe are clearly not representational items. In a nutshell, if representational content is underwritten by causal relations, then because causal relations are ubiquitous, so too is representational content. Or, if representational status reduces to an item being in a causal relation to another item, then virtually everything is a representation bearer.

The problem is to all intents and purposes identical to the one confronting the resemblance theory just discussed. Any attempt to construct content out of a type of objective relation confronts the existence of objective relations *without representation.* Like the symmetry-asymmetry objection, the problem of pansemanticism is a product of failing to distinguish between the question of representation and the question of content determination. And again like the symmetry-asymmetry objection, its straightforward resolution lies with distinguishing between these two issues. Representational content is not ubiquitous because causal relations are ubiquitous. Causal relations are content-conferring when the conditions for representational status are satisfied. The question of representation is a separate issue to the question of content determination, and the causal theory, an answer to the latter, *slots in* to an answer to the former.[34]

In this chapter I've sought to substantiate and clarify representationalism, as well as put the (admittedly skeletal, at this stage) resemblance theory on firm footing. The way is now clear to develop the resemblance theory in greater detail. As indicated in the Introduction, I'm going to arrive back at the resemblance theory via a path through the problem of misrepresentation. And I set out on such a path with an examination of the "orthodox approach" to content determination and misrepresentation, the topic of the next chapter.

[34] Fodor (1990a, 93) also offers a response to the problem of pansemanticism, but it differs from the one developed here. Fodor has particular difficulty in overcoming the pansemanticism objection because he operates with a *dyadic* conception of representation (as opposed to the triadic conception I've put forward above).

CHAPTER TWO

CONTENT DETERMINATION AND MISREPRESENTATION: THE ORTHODOX APPROACH

No one problem has haunted attempts to answer the question of content determination more than the difficulty of accommodating misrepresentation. The orthodox proposals for admitting misrepresentation, as is well known, confront serious difficulties. In this chapter I'll critically examine the orthodox approach to content determination and the problem of misrepresentation. The focus will be on the work of the most influential of recent theorists, namely, Fodor, Dretske, and Millikan. I'll detail the TsCD they develop, their proposals for accommodating misrepresentation, and the standard objections such proposals encounter. The overall goal of such a critical examination is to motivate the need for a new approach to content determination and misrepresentation. To this end, I aim to demonstrate that despite the differences between the orthodox TsCD and ways of catering for misrepresentation, there is a common, general type of strategy, or approach, adopted. The hope is that such a demonstration serves as something of a diagnosis of the difficulties recently encountered in accommodating misrepresentation. Concomitantly, my intention is to clarify the standard methodology in order to reveal the shape of an alternative approach to content determination and misrepresentation. The investigation begins with a brief account of the phenomenon of misrepresentation, and a look at how the crude versions of two TsCD render misrepresentation impossible.

1. The Problem of Misrepresentation

It is a fact that mental states sometimes (perhaps even reliably for all we *know*) misrepresent the world. The world as it is represented by mental states (the world mental content depicts) may not correspond to the world itself (the world as it is independent of any representation). Mental states can fail to be veridical. We are all familiar with various forms of visual illusions. We see mirages in the distance, straight sticks bent at the point of entering water, and coloured afterimages. To use the by now well-worn example, I might have a perceptual experience of a horse occasioned by the presence of a cow (Fodor 1987, Chp. 4). An animal in the wild might suddenly flee, having registered "danger" or "predator" at a rustle in the bushes caused by the wind. An unusual combination of various sounds may elicit an experience of the phone ringing, yet we soon find that we are mistaken. We can hear voices when no one is there, or hear someone say something other than what was actually said. We are capable of erroneous beliefs about a current state of affairs, and inaccurate recollections of events past.

In the most radical cases of misrepresentation, a mental state will represent an object as being present (and as being a certain way), when no such object is present. (Macbeth was notoriously prone to difficulties of this sort). A mental state also misrepresents if it represents an object as having certain properties and the object itself doesn't possess those properties (see Dretske 1995, 27). Third, a mental representation bearer misrepresents when the "comment" it makes of a feature, property, or variable – the "value" it attaches to a variable – does not correspond to the true situation. Actually, I'm inclined to view the latter not as a separate "kind" of misrepresentation, but as a useful way of thinking about the phenomenon of misrepresentation in general. For one thing, this "third type" of misrepresentation will subsume the other "two types" if the variable in question is a rather coarsely individuated one (for example, the local spatial layout, or the properties possessed by a certain object). For another thing, I'm going to subsequently recommend a rather fine-grained conception of mental content, one which I'll often talk about in terms of the "value of a variable".

A useful way of characterising misrepresentation – one which will be of central importance over the next few chapters – is due to Fodor (it comes from an in-conversation remark reported by Cummins 1996, 8): "Error occurs when a representation is *applied to* something it is not *true of*, for example, when one applies a representation of a cow to a horse." Cummins calls that to which a representation bearer is applied the bearer's *target*, and I'll often refer to it as the *intentional object*. What a representation bearer is true of corresponds to its content. Misrepresentation, then, is a mismatch between the content of a representation-bearer token and the actual state of affairs in the represented world to which the cognitive system applies the representation bearer. I'll have much more to say on the distinction between content and application over the course of the next few chapters.

Any theory of mental content determination must allow for the fact of misrepresentation. It's important to stress that accommodating misrepresentation is an ineliminable and central *desideratum* of a satisfactory TCD, and not merely a technical difficulty, a niggling peripheral issue to be ironed-out once one's TCD has been finalised. The capacity for misrepresentation is *constitutive* of the capacity for mental representation (see, for example, Dretske 1988, 65; 1995, 28; Lloyd 1989, 13). A TCD with no room for misrepresentation is not a viable contender.

My own perspective on the recent history of the problem of misrepresentation is that the standard proposals, in concert with their various TsCD, have a somewhat unwieldly, even *ad hoc*, air about them. Sometimes a purported solution to the problem of misrepresentation has the appearance of being "tacked-on" to a TCD. And in certain cases, either the proposal for misrepresentation, or the TCD itself, comes across as overly "theoretical", a product of philosophical fancy footwork, and lacking in independent support. Here my sentiments are with Cummins (1996, 5) who refers to the orthodox proposals as "precious": a little prodding and the cracks soon appear. As foreshadowed in the Introduction, I propose to adopt an alternative strategy in regard to the problem of accommodating misrepresentation. Rather than beginning with the determination of veridical content and then seeking to make room for misrepresentation – as I'll subsequently show, this is an central element of the orthodox approach – I'm aiming to build-in the capacity for misrepresentation from the outset. That is, I want to approach the determination of content in a manner which is neutral between veridical and non-veridical content. Whether content is true or in error, on this approach, is an issue *beyond* content determination. I'll arrive at such an alternative strategy by first examining the orthodox approach, and isolating its core commitments and assumptions. A new approach will take shape as these assumptions are subsequently relinquished. But I'm getting somewhat ahead of matters. The first task is to show why misrepresentation is such a difficulty for attempts to construct content out of some kind of objective relation between representation bearers and representational objects. I'll begin by showing how the crude versions of two TsCD render misrepresentation impossible. The first example is the traditional resemblance theory.

According to the traditional resemblance theory, content determination is a matter of a first-order isomorphism between representation bearer and representational object, meaning that bearer and object have physical properties in common. That is, a representation bearer has as its content that environmental object it resembles (in some specified respect), or perhaps resembles most closely, in terms of its intrinsic physical properties. Such a simple resemblance theory, however, faces a problem in admitting misrepresentation. If representing a certain object is a matter of a representation bearer resembling that object, then misrepresenting that object would seemingly consist in "misresembling" it (not resembling it, or perhaps, not resembling it in certain respects). However, this cannot work. For one thing, on this idea the difference between misrepresenting something, and not representing it at all, is difficult to specify, and maybe even breaks down completely, which clearly won't do. More importantly, strictly speaking, on the resemblance theory there is *no sense* to the idea that a representation bearer fails to resemble that which it represents. The theory cannot both say that a representation bearer has environmental object *E* as its content by virtue of its resemblance to *E*, and yet fails to resemble *E*.

Take a case of radical misrepresentation: Macbeth's hallucination of a dagger. The resemblance theory cannot make sense of such a case. According to the theory, Macbeth's mental representation bearer resembles a dagger before him, so its representational object is a dagger before him. Macbeth's perceptual experience cannot both resemble a dagger and fail to represent a dagger; the experience cannot, that is, both resemble (and thereby represent) a dagger and yet misrepresent it. Mental misrepresentation is a fact, but the (crude) resemblance theory cannot both explain content and allow for misrepresentation. The resemblance theory, then, at least in the form presented here, won't do for mental content determination.

Consider now Jerry Fodor's version of the causal theory. Fodor's account is developed in the context of his Representational Theory of Mind. Accordingly, it is the primitive nonlogical symbolic representation bearers of the Language of Thought (or "mentalese") which are at the "bottom of the pile" (1987, 98) when it comes to the question of content determination.[35] Fodor's first pass at an answer – the "Crude Causal Theory" – is that mentalese symbol-types have their contents by virtue of certain causal relations which obtain between symbol tokenings and property instantiations in the external world. The content of a mentalese symbol is

[35] Fodor has defended his causal theory over a number of works; see his 1987, 1990a, and 1994. For the most part, I'll focus on his 1987 presentation.

that property which causes the symbol to be tokened. In Fodor's words: "...symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings" (1987, 99). The causal relations holding between symbol types and the properties responsible for their tokenings are, for Fodor, counterfactual supporting. That is, it's a *law* that instances of the property *horse*, for example, cause tokenings of the symbol type HORSE, or instances of the property *horse* would cause tokenings of the symbol type HORSE were such instantiations to occur (1987, 99). In short, a primitive symbol type has as its content the property with whose instantiations it *reliably covaries*.

As Fodor (1987, 101-102) acknowledges, the Crude Causal Theory (CCT) renders misrepresentation impossible. The fact is that a given type of mentalese symbol – such as HORSE – reliably covaries with properties other than *horse*. *Cows-on-a-dark-night*, for example, reliably cause HORSE tokenings. Such tokenings are misrepresentations since, by hypothesis, the symbol type HORSE expresses the property *horse*, not *cow-on-a-dark-night*. But the CCT does not mark them as misrepresentational tokens. For, according to the CCT, the content of a symbol is that property with which it reliably covaries. So, on the CCT, a symbol whose tokening is caused by *horse* or *cow* has as its content not "horse" (or "cow") but "horse-or-cow". And if HORSE expresses the disjunctive property "horse-or-cow" then *cow*-caused HORSE tokenings are not misrepresentational. So, the causal TCD, at least in its crude form, has no room to admit misrepresentation.

Original-intentionality TsCD attempt to build content out of a certain kind of objective relation between representation bearers and representational objects. However, the one, as it were, "unbroken" relation cannot serve the dual roles of determining content and yielding instances of misrepresentation. The objective relation secures veridical content, and there is just no room to manoeuvre misrepresentation into the picture.

2.2. The Orthodox Approach

The aim of this section is to show that despite the differences between certain current TsCD and proposals for accommodating misrepresentation, there is the adoption of a common approach to content determination and misrepresentation. I will first set-out the orthodox approach in general terms, and then go on to show how it is embodied in the particular theories of Fodor, Dretske, and Millikan.

The orthodox approach, I submit, can be characterised in terms of the following commitments or assumptions:

> (1) *Begin* with the determination of *veridical* content. The starting-point of a TCD, that is, is concerned with the contents of mental states which are true of the world.

It is here that we find the roots of the orthodox methodology. The standard approach is to begin by constructing a TCD in such a way that content is veridical and misrepresentation impossible. The theorist *then* has to find a way of bringing in the possibility of misrepresentation. This involves an amendment to, or extension of, the original story concerning content determination. The recent strategy has been to specify a certain *privileged subclass* of circumstance with respect to, or within, the objective representation bearer-representational object relation. The special circumstance is deemed as providing for content determination. Token representation bearers produced in *other* circumstances are then misrepresentations.

> (2) Admitting misrepresentation consists in finding a *difference* between the circumstances which determine the content of representation bearers of a certain type, and circumstances where a token representation bearer misrepresents. The latter circumstances – circumstances which (purportedly) admit misrepresentation – are not content-conferring.

The account of misrepresentation, that is, specifies how conditions are different to those which determine the content of representation-bearer types.[36] On the orthodox approach, then, the full-blown TCD itself includes a specification of the condition which determines content, and a specification of a different condition which is not content-conferring and constitutes the account of misrepresentation.

> (3) The accommodation of misrepresentation, and providing an account of misrepresentation, are required elements of the TCD itself.

That is, a statement of misrepresentation must be included *within* the account of content determination.

---

[36] This element of the orthodox strategy will become clearer when I discuss particular instances below.

I'll now examine how the above commitments manifest themselves in the work of Fodor, Dretske, and Millikan. I'll also be putting forward certain central problems which arise for these theorists' attempts to incorporate the capacity for misrepresentation.

2.1. Fodor

Fodor's TCD, as described above, appeals to counterfactual-supporting (that is, nomic) causal relations between mental representation bearers (primitive mentalese symbols) and the environmental conditions (properties) they represent. The content of a given symbol type is that property which, were it instantiated, would cause a tokening of the symbol. The contents of symbols, on this story, are "ipso facto veridical". That is, "the condition for an A-tokening meaning 'A' is identical to the condition for such a token being true" (Fodor 1987, 101). Misrepresentation is thereby ruled-out. But the occurrence of misrepresentation is a fact. So Fodor has to find a way of letting-in misrepresentation.

What an advocate of the causal theory wants in order to accommodate misrepresentation is that a symbol type expresses a certain property (*A*), and that a token of the type is produced by a property other than that the type expresses (*B*). That is, *B* brings about an A-token *without* contributing to the content of the A-token. The CCT, however, won't admit such a situation. Again, according to the CCT, the content of a symbol type is that property whose instantiations reliably bring about tokenings of the symbol. If *B*-instantiations, as well as *A*-instantiations, can cause tokenings of A, then the most reliable connection is between *A-or-B*-instantiations and A-tokens. A's content is then not "A", or "B", but "A-or-B". And if A expresses the disjunctive property "A-or-B", then *B*-caused A-tokenings are not misrepresentations (Fodor 1987, 101-102). Fodor calls this the "disjunction problem". Given, as defenders of the causal theory acknowledge, that any mental representation bearer is apt to be tokened by a variety of conditions, a solution to the problem of disjunctive content, such that misrepresentation is admitted, must be found. More generally, and as Fodor (1990a, 91) has stressed, a viable TCD must respect the *robustness* of content. The thinking is that any representation bearer has discrete (nondisjunctive) content despite the heterogeneity of its causes.

To deal with the disjunction problem – to allow for misrepresentation – the CCT must be amended so that not all the causes of a symbol tokening confer content to the symbol type. Only the "true"-cause-to-mentalese-token relation is to determine content. As suggested, the task is one of finding a difference, expressible in nonrepresentational terms, between *A*-caused A-tokenings and *B*-caused A-tokenings (see Fodor 1987, 106). That is, the task is to say why symbol tokens (of any given type) caused by instances of properties other than the property which, by hypothesis, determines content are misrepresentations (or "wild" as Fodor calls them).

Fodor's proposal turns on the *asymmetric dependence* of certain causal relations on others. There is, he claims, a pivotal asymmetry between the "true" and "false" causes of a symbol tokening. The false causes of a symbol tokening – those which don't contribute content to the symbol type – are nomically dependent on the true causes, but the converse dependency does not hold (1987, 107-111; 1990a, 91-92). This means that severing the true-cause-to-mentalese-token link severs the false-cause-to-mentalese-token link, but severing the false link does not thereby sever the true. The false causes of a symbol tokening, to put it somewhat metaphorically, "hitch a ride" on the nomic connection which obtains between a symbol and its true cause. (Were there no drivers there wouldn't be hitch-hikers, but if there were no hitch-hikers there would still be drivers, to put it rather a lot metaphorically). So, *cow* wouldn't cause HORSE were it not for the *independent* existence of a *horse* to HORSE nomic connection. The true-cause-to-symbol connection is privileged; it has, so to speak, ontological priority. Fodor's appeal to asymmetric dependence provides a necessary condition for *B*-caused A tokens to be instances of misrepresentation (wild), and his full-blown TCD incorporates this condition (see 1987, 109; 1990a, 121).

It can now be seen how the commitments or characteristics of the orthodox approach outlined at the beginning of this section are present in Fodor's theory. First, the idea is to begin by securing veridical content, and then attempt to incorporate misrepresentation. Second, because the starting point is the TCD's securing of veridical content, incorporating misrepresentation involves stating a difference between the circumstance which determines symbol-type content and the circumstance which obtains in cases of misrepresentation. For Fodor, as just shown, the circumstance is different in cases of misrepresentation because the causal connection to wild tokenings is asymmetrically dependent on the causal connection to well-behaved ones. Finally, and again because the foundation of the TCD is an account of veridical content, the admission and account of misrepresentation – the specification of the non-content-conferring circumstance – is a component of the TCD itself (see Fodor 1990a, 121 where this is particularly stark). The following passage more or less expresses this third characteristic of the orthodox approach:

> ...we need a difference between A-caused 'A' tokenings and B-caused 'A' tokenings that can be expressed in terms of nonintentional and nonsemantic properties of causal relations; for nonintentional and nonsemantic properties of *causal relations are all that the Crude Causal Theory of Content has to play with* (Fodor 1987, 106, emphasis added).

As suggested, on the alternative approach to content determination and misrepresentation to be subsequently developed, the truth or falsity of representational content is an issue which lies outside of the TCD. Resources distinct from those employed in the account of content determination, that is, are utlilised to accommodate misrepresentation.

It's fair to say that Fodor's proposal for misrepresentation has not exactly been greeted with widespread enthusiasm (something Fodor himself has recognised; 1990a, 89). Naturally enough, many of the criticisms of Fodor's idea focus on whether there really is an asymmetric dependence relation between the true-cause-to-symbol-tokening connnection and the false-cause-to-symbol-tokening connection (see, for example, Cummins 1989, Chp. 5; Cummins 1996, Chp. 5; Godfrey-Smith 1989; Maloney 1990; Manfredi and Summerfield 1991; and Sterelny 1990). One of the most straightforward, and to my mind most telling, objections to Fodor's proposal goes as follows. We are simply not given a reason to believe that Fodor is entitled to assume an *independent* and *privileged* causal set-up between the "true" causes of a symbol tokening and tokenings of that symbol type (between, for example, *horse* and HORSE). Fodor is simply *stipulating*, and, to be blunt, begging the disjunction problem. Fodor stipulates that it is the *horse*-to-HORSE causal relation which is privileged (in contrast, for example, to the *cow-or-horse*-to-HORSE causal relation). And it may very well be thought that it's his theory and he gets to stipulate what the theory asserts. However, it's highly questionable that Fodor is entitled to privilege the sort of causal relations he does when, for example, it is not *horse* but the "larger, unnatural kind *K*...which includes typical horses, slim cows, muddy zebras, and other things apt to be mistaken for cows" which is "the most basic causal relation" (Godfrey-Smith 1989, 539). That is, HORSE covaries *more reliably* with *K* than it does with *horse*. And it seems plausible to suppose that were it not for the correlation between *K* and HORSE, neither cows nor horses would cause HORSE tokenings.

The causal theory seeks to construct content out of lawful correlation. But it is the link between *K* and HORSE which clearly looks to be the most robust nomic relation. Fodor's claim might be that it is the causal relation between the *property* of being a horse and HORSE which is relevant (indeed, perhaps the "language of thought" requires it to be so). The causal link between those things which look like horses ("horsey looks") and HORSE is beside the point. However, is it plausible to suppose that an objective relation between a (natural kind) property (*horse*, for example) and symbol tokenings is privileged or more basic, while one between the (unnatural) kind of "horsey looks" is parasitic? Properties might seem more ontologically robust than appearances or "looks" to philosophers. But what we're concerned with here is how Nature cobbled together systems capable of representing (and misrepresenting) their environments, and Nature is going to take the most rough and ready, or economical, route, not necessarily the metaphysically cleanest.

Cummins (1989, 58) raises a very similar objection to that brought by Godfrey-Smith. Changing the example, consider again what is required for the asymmetric dependence proposal to work:

(1) If it were the case that mice didn't bring about MOUSE, then it would be that shrews didn't bring about MOUSE.

(2) If it were the case that shrews didn't bring about MOUSE, then it would be that mice didn't bring about MOUSE.

For asymmetric dependence to admit misrepresentation, (1) must be true and (2) false. However, can this be the case? Assume that shrews bring about MOUSE because they look like mice. So, if it were the case that shrews didn't occasion MOUSE, then that might be due to the fact that (a) shrews didn't look like mice, or (b) mouse-looks didn't bring about MOUSE. If the latter, however, neither would mice bring about MOUSE, and so (2) would be true, contrary to what's required for the asymmetric dependence proposal.

An extension of the above line of argument goes as follows (Cummins 1989, 60-61). Consider again (2). There are two ways that shrews don't cause MOUSE: i) Mousey looks don't occasion MOUSE. But if this were the case, mice wouldn't occasion MOUSE either. So, it must be that, (ii) shrews don't cause mousey looks. But mice have mousy looks, and so occasion MOUSE. Thus, as required, (2) is false.

Now examine (1). There are two ways that mice don't occasion MOUSE: (iii) Mousey looks don't occasion MOUSE. So shrews, looking like mice, won't occasion MOUSE either (by "hitching" on mousey looks), and (1) is true, as demanded. But, if this is so, (2) is not, as required, false, but true. It must then be that (iv) mice don't cause mousey looks. Maybe, as Cummins suggests, mice die-out, or for some reason their appearance changes. The shrew-to-MOUSE link, however, remains intact, and so (1) is false, contrary to what's needed.

So, we have to apply the "shrews don't cause mousey looks" clause (that is, (ii)), in order to secure the falsity of (2). But we need to utilise the "mousey looks don't cause MOUSE" clause (that is, (iii)), in order to secure the truth of (1). There is no consistent interpretation that establishes the truth of (1) and the falsity of (2) (Cummins 1989, 61).

As Cummins admits, a determined defender of asymmetric dependence might very well claim that the account of asymmetric dependence *is* so convoluted: (1) and (2) do cash-out in different ways. But, and here's what I think really bugs critics of Fodor's proposal, if things need to get this complex or cumbersome, then do we really have a natural means of underwriting the distinction between veridical and false content? One is entitled to think that Fodor's proposal is all very clever and the like, but as with all good sleight-of-hand tricks, we know that things couldn't really have happened the way the magician wants us to believe.

The above discussion looks to only a couple of considerations concerning an attempt to solve the problem of misrepresentation which has attracted a great deal of philosophical scrutiny. But it is not my aim here to provide a comprehensive survey of the debate. As indicated, I'm more interested in isolating the characteristics of the orthodox approach, and providing some reason for thinking that a fresh approach is worth investigating. Let me turn, then, to the work of Fred Dretske.

For almost two decades now, Dretske has grappled mightily with the task of constructing a TCD able to admit of misrepresentation. Over this period, he has put forward two distinct proposals. The first was developed in his influential *Knowledge and the Flow of Information* (1981). The second was initiated in his paper 'Misrepresentation' (1986/1990a), and developed further in *Explaining Behavior* (1988) and *Naturalizing the Mind* (1995). My discussion of Dretske's work will consequently be divided into two parts.

2.2. Dretske the First

Dretske's earlier work was a pioneering effort in the development of causal approaches to content determination. Dretske (1981) glosses representational content in terms of "information", and thus content determination as the conferral of information upon internal representation bearers. Content determination is a matter of information-bearing signals – generated by states of affairs in the external world – causing the tokening of mental representation bearers. His approach is therefore often referred to as "informational semantics". Dretske's specialised treatment of the notion of information, and the presentation of his theory, is somewhat detailed. But we need not be concerned with the technicalities here. At the heart of his theory is the idea of natural indication, or counterfactual-supporting correlation. The foundation for other causal theories (most notably Fodor's version) is thereby clearly present in Dretske's work. In essence, a type of mental item bears information about – has as its content – a type of environmental object insofar as instances of that object are (nomically) causally responsible for tokenings of the mental item.

As illustrated in the discussion of Fodor's causal theory, however, there is a problem in providing for misrepresentation. What Dretske wants is that a mental representation bearer *M* represents, for example, environmental object *E* in cases where *M* is caused by *F*. But the causal theorist cannot coherently assert this because causal correlation confers content, and so *F* is in *M*'s extension. A token mental representation bearer (*M*) of a given type is causally correlated with a number of environmental objects (*E, F*). The correlation of *M* with *E-or-F* is more robust than the correlation of either *M* and *E* or *M* and *F*. Since causal correlation makes for content determination, *M* is a reliable indicator of not *E* (or *F*) but the disjunction *E-or-F*. An *F*-caused *M*-token is therefore not a misrepresentation of *F* as *E*, since *M* represents *E-or-F*.

Dretske's first proposal for dealing with misrepresentation sees him employing the notion of a *learning period*, marking a sharp distinction between this period and the time subsequent. The content of a representation bearer type is determined during the learning period. During this time conditions are optimal in the sense that a "teacher" serves to establish a correlation between the representation bearer type and a certain representational object; the content of the representation bearer is thus the representational object with which it is so correlated. *After* the learning period, however, conditions are no longer ideal in that the rigorous enforcement of the content-determining correlation is no longer in place. During this time, the representation bearer tokens may be brought about by something *other* than the representational object. Such tokens are thereby misrepresentations.

Dretske's account clearly exemplifies the orthodox approach to content determination and misrepresentation. He begins with veridical content, the determination of which is explicated in terms of nomic causal correlation. The basic TCD must then be augmented with an account of misrepresentation. This account is, in effect, part of the TCD itself. That is, the augmented TCD specifies that content is determined in some sort of special circumstance; a special subclass of the representation-bearer-to-representational-object relation is privileged. For Dretske, as just described, this special circumstance is the learning situation. When the situation is other than that of the special (content-determining) circumstance – for Dretske, post learning period – misrepresentation is free to occur. Again, at the heart of the orthodox approach is a distinction between a certain kind of circumstance under which content is determined (and content is veridical), and circumstances which are non-content-conferring (those which yield the capacity for misrepresentation). Such a characterisation is not particularly radical. Fodor (1990a, 60, 64, 89-90) has described the strategy adopted by Dretske (and Millikan) as one of distinguishing between two types of situations. In "*type one*" situations, causes of representation-bearer tokenings are content-conferring. This is the type of situation "in which the tokening of a symbol can be caused only by things that are in its extension...." (Fodor 1991, 272). In type one situations, representational content is *ipso facto* veridical. In "*type two*" situations, the causes of representation-bearer tokenings are not content-conferring. In this type of situation, "symbols are allowed to be caused by things that they don't [represent]" (Fodor 1990, 61). Token representation bearers in type two situations are thereby free to be misrepresentational.[37]

Dretske's proposal, as Fodor (1984/1990a, 1987) has highlighted, is beset by a number of difficulties. For one thing, the distinction between the learning situation and the period after isn't sufficiently principled to yield a distinction between veridical and false content (Fodor 1984/1990a, 41). By what means does a representation bearer's content stop being shaped and suddenly become robust? Second, and even more significantly, Dretske's proposal is not internally coherent. Consider what Dretske is claiming. During the learning period a "teacher" ensures that the cognitive agent produces representation-bearer tokenings *R* only in response to environmental object *E*. Post learning period (whatever, precisely, is supposed to demarcate this latter period from the learning situation), the cognitive agent has a *R*-token occasioned by something other than *E*, say *F*. Since the *F*-caused tokening of *R* occurred outside of class, *R* (having *E* as its content) falsely represents *F* as *E*.

The problem with this manoeuvre, however, is that it conveniently but erroneously neglects to take into account the idea that if *F* is able to occasion *R*-tokens after the learning period, then *F* is capable of bringing about *R*-tokens *during* the learning period. Clearly, then, the causal correlation forged during the learning period is between *R* and *E-or-F*, and not, as Dretske requires, between *R* and *E*. The content of *R* is thereby *E-or-F*, and since an *F*-occasioned *R*-token is veridical when *R*'s content is *E-or-F*, Dretske hasn't escaped the disjunction problem.

As Fodor (1990a, 42) notes, Dretske has a possible reply available here. Dretske could claim that were an errant (*F*-caused) *R*-token to be produced during training, the teacher would *correct* the response. *R*-tokens represent *E* (and not *E-or-F*) because the correlation between *R* and *E* is the disciplined one. After training,

[37] Fodor (1990a, 91) claims that his proposal for misrepresentation doesn't fit into the type one/type two schema as defined above. Strictly speaking, on the precise definition of type one and type two situations given above, that may very well be so. Nonetheless, my claim is that Fodor's attempted solution still exhibits the general pattern of explanation exemplified by appeal to the distinction between two kinds of conditions or circumstances. That is, Fodor's proposal conforms to the orthodox strategy of isolating a privileged subclass of representation-bearer-to-representational-object relation as the content-determining circumstance.

and, by hypothesis, in the absence to the teacher, errant R-tokenings would go uncorrected and thereby constitute errors.

But this reply has problems of its own. First, the proposal is no longer naturalistic in that it appeals to the teacher's intentions, themselves intentional states, as the essential content-determining element. As Godfrey-Smith (1989, 538) suggests, one could claim that it is the teacher's *behaviour* that matters, rather than the teacher's intentions, so the threat of circularity can be kept at bay. Moreover, conceiving of the teacher as an intentional agent is not necessary; any type of "feedback" mechanism might do the job. Perhaps even the organism's own behaviour, I would claim, could play the required role. In this sense, the environment would serve as a "teacher". An organism, having tokened R in response to F, would find its behaviour unrewarded, or in some way "inappropriate". But it's not clear that the feedback here would be sufficiently selective, able to provide for the required discrimination between two or more environmental objects which, after all, may be very much alike. Moreover, the suggestion seems to apply only to to rather sophisticated representational creatures (those capable of closely and accurately monitoring their own performance).

Second, however, and even if the violation of naturalism could be avoided, it's clear that the response being considered here constitutes a radical departure from the original content-determination story. Rather than content determination being a matter of causal covariation – rather than appeal to a kind of objective relation between representation bearers and representational objects – the essential content-determining work is being done by the *effects* of representation-bearer tokenings. Though it goes unremarked, this, to my mind, is to abandon the idea that content is causally efficacious, and Dretske has always been a firm advocate of the view that a satisfactory TCD must accord with the causal potency of content. (This point will be revisited, and further explained, below). He therefore should reject the response countenanced here.

Dretske's original way of attempting to handle misrepresentation, then, is rather unpromising. Dretske himself has been moved to develop a different account, and it is to this more recent proposal that I now turn.

2.3. Dretske the Second

The TCD Dretske currently favours has two components. The foundation remains causal correlation (or "indication" as it is called in Dretske's discussion), which is augmented by appeal to teleological considerations. There is, then, a conspicuous shift in emphasis in Dretske's more recent attempts to construct a workable TCD. Rather than unpacking content determination solely in terms of counterfactual-supporting correlation, Dretske employs the idea of "natural functions" – which, as I'll show, he seeks to secure by appeal to an organism's behaviour – and the corresponding idea of "functional meaning". The significant change in the character of Dretske's approach to content determination only serves to highlight just how defiantly the problem of misrepresentation has resisted resolution. In a very real sense, the problem of misrepresentation has driven the direction of Dretske's thinking on content determination. I begin with a very brief look at the indicator component of Dretske's theory.

Indication underwrites what Dretske (following Grice) calls an item or state's "natural meaning". A state involved in an indication relation is a "natural sign" (1986/1990a). However, mental content determination cannot simply be a matter of indication because, as Dretske recognises, natural meaning does not admit misrepresentation. Dretske (1986/1990a, 131, citing Grice) writes: "...an occurrence (a tokening of some natural sign) means (in...the natural sense of "meaning"...) that *P* only if *P*." A state's natural meaning is always (of necessity) veridical. If a natural sign (a sign of *m*) is tokened when what it is a sign of does not obtain (*n*), then it does not misrepresent *n* as *m*. It only means *m* when *m* actually obtains.

We are, of course, in familiar territory here. Unlike natural signs, Dretske acknowledges, as he must, that any given mental representation bearer may be reliably correlated with – and thereby indicate – a range of environmental objects. In such cases, the most reliable correlation is between a given type of mental representation bearer and the (probably unnatural) kind *K* (where *K*, of course, is the set of environmental objects whose individual members are each sufficient to occasion a tokening of the representation-bearer type). Thus indication secures only disjunctive content; we're back with the disjunction problem.

With, then, at least one eye to the problem of accommodating misrepresentation, Dretske proposes to pick-out a particular environmental cause of a bearer's tokening, identifying it as the bearer's content, by appeal to teleological considerations. His full-fledged TCD, as well as his account of misrepresentation, relies on the notion of "*functional meaning*". A state's functional meaning – its content – is the environmental object or condition it has the *function* of indicating (1988, 59, 70, 77; 1995, 2). Though a given state (*C*) may be tokened by various environmental conditions, there is only one thing *C* is *supposed* to indicate, namely, that condition the indication of which *C* was *selected for*.

Biological systems, Dretske claims, have natural indicator functions. They possess states and mechanisms[38] which are supposed to indicate the presence of, or provide information about the state of, certain environmental conditions. But Dretske is by no means out of the disjunctive woods yet. The idea of natural selective forces, baldly stated, is too blunt an instrument to confer upon a representation-bearer type the function of representing one environmental condition, out of all the conditions the bearer indicates (a point of which Fodor repeatedly reminds teleologists; see, for example, his 1990a, Chps. 3 and 4). Relatedly, the idea that organisms have natural indicator functions is controversial; more so, the idea that organisms (the representational systems they possess) have *determinate* natural functions. And it is the determinacy of natural function that Dretske must establish in order to solve the disjunction problem, as he well recognises.[39] Consequently, Dretske (1988) offers a certain specification of what it is that a mental representation-bearer type is selected for, a particular way of unpacking the claim that the representational systems organisms possess have

[38] Dretske equivocates on whether it is cognitive states themselves, or merely the mechanisms which produce such states, that have natural functions. There is potential for problems here. I'll return to this point towards the end of the current subsection. See also note 40 below.

[39] In what is surely one of the more unlikely instances of strange bedfellows, Dennett and Searle are against Dretske on the idea of naturally acquired functions, as Dretske (1995, 7) acknowledges. According to Searle (1992, 52), functions are "entirely in the mind of the observer". For Dennett (1987, 287-321), there are no natural functions because functional attributions are always subject to indeterminacy. The indeterminacy arises from Dennett's view that we must adopt the intentional stance toward Mother Nature, and the existence of rival, but equally explanatory, interpretations is a characteristic of taking-up the intentional stance.

natural indicator functions.[40] It is at this point that Dretske appeals to an organism's behaviour.

A state *C*'s natural indicator function, he claims, results from its recruitment as a cause of a specific behavioural response *M*. Response *M* is a characteristic behaviour caused by *C*'s tokening, a behaviour that is ecologically appropriate. According to Dretske, we can say that *C* has the function of indicating *F* because it is *C*'s indication of *F* which explains its recruitment as a cause of *M*. That is, *C*'s indication of *F* acquires explanatory significance in regard to the establishment of a causal link between *C* and *M*. In Dretske's words:

> Once C is recruited as a cause of M – and recruited as a cause of M *because of what it indicates about F* – C acquires, thereby, the function

---

[40] The talk of "natural indicator functions" here is meant to subsume both those (purported) indicator functions which are a product of evolutionary natural selection, and those conferred by the kind of selective forces operative in cases of individual learning (specifically, for Dretske (1988), operant or instrumental learning). While both kinds of selective forces are part of Dretske's account, in his 1988 work it is clear that the latter plays the primary role (see especially Chp. 4). There is a certain "looseness" of interpretation, and even confusion, about this point in the literature. Fodor (1990b), for example, presents his exposition of Dretske's theory solely in terms of evolutionary selective forces. Dretske acknowleges the potential confusion, and provides a note of clarification:

> In order to get meaning itself (and not just the structures that have meaning) to play an important role in the determination of an *individual's* behavior (as beliefs and desires do) one has to look at a meaning that was instrumental in shaping the behavior that is being explained. This occurs only during individual learning. Only then is the meaning of a structure type (the fact that it indicates so-and-so about the animal's surroundings) responsible for its recruitment as a control element the production of appropriate action. It is at this point, and not before, that belief (and other intentional attitudes) enter the explanatory picture. Natural selection gives us something quite different: reflex, instinct, tropisms, fixed-action-patterns, and other forms of involuntary behavior – behavior that is (typically) *not* explained in terms of the actor's beliefs and desires (if any). These genetically determined patterns of behavior often involve (as triggers for response) internal indicators (information-carrying elements), but, unlike belief, it isn't *their* content that explains the way they affect output. That is determined by the genes (1990b, 14-15).

Dretske's primary concern in his 1988 and 1990b is with how meaning – representational content – is implicated in the causation of behaviour. In order to address this issue he must, of course, answer the question of content determination. And it is in such a context that the issue of how natural indicator functions are conferred arises.

Dretske's position on evolutionary selective forces, and those operative in learning, at least insofar as I understand it, looks to be something like the following. For *genetically determined* behaviours, internal indicators have natural indicator functions by virtue of evolutionary selection. But their content is not relevant to an explanation of the behaviour they occasion. For *learned* behaviours, the "*structures* that carry or embody information" are there, and have the indicator functions they do, because of evolutionary selection. The "*information*" or content these structures carry – the information "put to work…in the production and control of behaviour" – is a product of an individual organism's learning history (Dretske 1988, 96). In the latter kind of case, content *is* relevant to an explanation of behaviour. In a later work, Dretske writes: "I will…assume in this work that natural indicator functions are always acquired through some historical process like natural selection (for systems) and learning (for states)" (1995, 170 note. 4). So, Dretske's view seems to be that, for non-genetically determined behaviours, an organism's *mechanisms*, structures, or systems for producing cognitive states are biologically endowed. The informational *content* of cognitive *states* – derived from their natural indicator functions – is a product of the organism's learning history.

> of indicating F. Hence C comes to *represent* F. C acquires its semantics, a genuine meaning, at the very moment when a component of its natural meaning (the fact that it indicates F) acquires explanatory relevance (1988, 84).

The above passage puts forward the centrepiece of Dretske's attempt to secure determinate indicator functions for mental-state types, and thereby the heart of his more recent proposal for mental content determination.

According to Dretske's theory, *F* is an environmental condition it would be useful for the organism to "know" (have information) about, and *M* is a behavioural response appropriate to the presence of *F*. Being appropriate, such behaviour facilitates the organism's survival. Representation bearer *C* is recruited as a cause of a behavioural response appropriate to the presence of one environmental condition, among many, *C* indicates, namely *F*, because the behaviour is appropriate to one thing *C* indicates, that is, *F*. Selective forces ensure that an organism attains a survival-enhancing co-ordination of its behaviour with its environment. Tokenings of mental states which cause a particular behaviour are produced by the presence of environmental conditions to which the behaviour is appropriate.

So, Dretske picks-out the content-determining cause of a state's tokenings in a teleological fashion. The content of a state (*C*) is that environmental condition (*F*) the indication of which led to the recruitment of *C* as a cause of behaviour appropriate to the presence of *F*. All the other causes of *C*'s tokenings (all the other conditions *C* indicates) are non-content-conferring. In this way Dretske proposes to allow for the possibility of *misrepresentation*. On Dretske's view, as pointed out above, biological systems have natural indicator functions: internal states and mechanisms are supposed to provide veridical information about the presence of, and features possessed by, certain environmental conditions. However, biological systems, like representational systems generally, are not perfect: sometimes a state will not *perform* its indicator function (either due to something going wrong with the system itself, or because conditions are less than optimal). Sometimes, that is, a state will be tokened by an environmental condition it is not the state's function to indicate. Still, Dretske claims, even when a system fails to perform its indicator function (when it doesn't indicate what it is supposed to indicate), it *retains* its indicator function (and thereby its functional meaning, its content).

Misrepresentation, Dretske thinks, is to be explained in terms of a system failing to indicate what it is supposed to indicate (1995, 2, 4, 28). In short, a state's content is determined in teleologically normal circumstances; misrepresentation occurs when something other than the condition the state is supposed to indicate causes the state's tokening.

Dretske's way of providing for the determination of content and the capacity for misrepresentation conforms to the by now familiar strategy. At the foundation is the determination of veridical content. Here Dretske appeals to indication, that which confers a representation bearer's natural meaning. As natural meaning doesn't enable misrepresentation, Dretske must augment the basic content-determining story. This he does by constructing the notion of functional meaning, and the contents of representational states are identified with such a notion. So we have here the standard gambit of marking a distinction between a kind of circumstance which is content-conferring (where the causes of representation-bearer tokenings are identified as the representation-bearer type's content), and those circumstances where the causes of bearer tokenings do not contribute to the type's content (such that bearers so caused are purportedly misrepresentations). When conditions are teleologically normal (the specified "type one" situation) – when, that is, cognitive states/mechanisms are performing in accordance with what they were selected for – a representation bearer is tokened only by that environmental condition it is the type's function to indicate; the environmental condition, that is, which selective forces have co-ordinated with its distinctive behavioural response. All other causes of the representation bearer tokenings are not teleologically normal (the "type two" situation): the tokenings of the bearer they occasion were not responsible for the bearer being recruited as a cause of its distinctive behavioural effect. Finally, the accommodation of misrepresentation is very much a part of the TCD itself. Dretske renders content determination in terms of the conferral of functional meaning, and the notion of functional meaning is expressly designed to let-in the capacity for misrepresentation. To put it in a slightly different way, the account of misrepresentation utilises the same basic resources as are employed in the TCD. Dretske adds a teleological component to indication so as to pick out the content-determining cause of a mental state, and thereby let-in the possibility of misrepresentation (a state being tokened by something other than its content-determining cause). The addition of a teleological component to indication constitutes Dretske's TCD, yielding the notion of a state-type's indicator function, its functional meaning. Misrepresentation is accounted for in terms of a state failing to perform its indicator function.

I turn now to some of the criticisms which have been leveled at Dretske's account of content determination and misrepresentation. While, as suggested above, the very character of his TCD has been shaped by the problem of misrepresentation, ultimately I don't think Dretske's proposed solution is successful. Moreover, I believe his struggle to overcome the problem of representational error has resulted in a TCD which admits of certain serious defects. And while I'm more interested in the problems with his proposal for misrepresentation, I will consider a difficulty which pertains (at least initially) to the TCD itself (though, as suggested, on the orthodox approach the TCD and account of misrepresentation are bound-up together). As I will show, such a difficulty carries through to undermine Dretske's purported solution to the disjunction problem.

According to Dretske (1988), a state *C* acquires the function of indicating (just) one environmental condition, *F*, through its explanatory role: the indication relation (holding between *C* and *F*) enters into the explanation of why *C* causes a certain behaviour, *M*. The forging of the causal link between *C* and *M* is explained by the indication relation between *C* and *F*. *M*, the characteristic behaviour *C* brings about, is appropriate to the presence of *F*. So, according to Dretske, *C* has the natural function of indicating *F*.

The problem with this line of reasoning, as Fodor (1990b, 33) nicely puts it, is that it "buys the specificity of the state's environmental *cause* by just *assuming* the specificity of its behavioural *effect*." As suggested, what Dretske needs to show is which of the indication relations that obtain between *C* and various environmental conditions is the content-determining relation. Dretske's proposal is that it's the one that selective forces have correlated with a characteristic behaviour resulting from *C*'s tokening. The difficulty is that tokenings of particular mental-state types *don't have characteristic behavioural effects*. (Except, of course, reflexes. But what we're concerned with here is intelligent behaviour). It's part of the very notion of *intelligent* behaviour that a certain mental state can produce any one of a whole range of behavioural responses; intelligence is the capacity for behavioural flexibility. Moreover, between the presence (and registration of) certain environmental conditions and the output of certain behaviours exists a great deal of cognitive machinery. An organism's behaviour toward its environment is mediated by a whole host of cognitive states. A given cognitive state may interact with any number of other cognitive states. And the production of any given kind of behavioural response is typically due to the activity of a number of cognitive states the organism is in and dispositions it possesses. Indication appears too early in the complex causal chain from environmental condition to behavioural responses in order for an internal state involved in an indication relation to be plausibly claimed as occasioning a characteristic behavioural effect. It is just not correct to claim, then, as Dretske does, that we can secure *C*'s indicator function by appeal to *C*'s characteristic behavioural effect. And, of course, if we can't secure a determinate indicator function for *C*, then the basis of Dretske's solution to the disjunction problem collapses.

The second criticism I want to bring against Dretske is that his account of content determination is ultimately in tension with his aim to operate within the view that content is causally potent. This problem spills over into his attempt to solve the disjunction problem. Dretske's commitment to the causal efficacy of content is apparent throughout his work. Indeed, his *Explaining Behavior: Reasons in a World of Causes* (1988) is a sustained attempt to defend the idea that it is *reasons* – the representational contents of mental states – which cause us to behave in the way we do. (For particular pronouncements, see the Preface and 79-80. See also 1986/1990a, 130).

Dretske's (1988) initial presentation of his theory vacillates between two different ideas. There is an equivocation on two different ways of explaining the forging of a causal link between *C* and *M*. And this means he equivocates on what, exactly, he is proposing as the content-determining factor. Dretske writes of a mental state indicating an environmental condition *or* having the *function* of indicating an environmental condition:

> ...[the] causal relationship between C and M, if it is going to be explained by something like the meaning of C, will have to be explained by the fact that *C indicates, or has the function of indicating*, how things stand elsewhere in the world (1988, 84, emphasis added; see also 79).

Here we find the root of the equivocation. Dretske is claiming that *C*'s content is determined by *C*'s indication relation, or by *C*'s indicator function. But these two are not equivalent, and Dretske cannot vacillate between the two. If *C*'s content is determined by virtue of its *indication* of *F*, then content determination is a matter of a causal relationship between representation bearer and representational object. And this constitutes a commitment to a TCD whereby content has a causal influence over the course of cognitive processing. The view that content is causally efficacious is essentially the idea that it is a representation-bearer's content which dictates its causal (cognitive and behavioural) effects. If, however, *C*'s content is determined by virtue of *C*'s *indicator function*, then content determination is a matter of a representation bearer's *effects*. Content, that is, is a consequence, not determinant, of cognitive processing. Dretske cannot have it both ways. Either mental content determines cognitive role, or cognitve role determines content. The equivocation is a serious mark against his TCD.

In fact, however, when Dretske comes to the definitive statement of his theory, it is clear that he thinks cognitive role is the crucial content-determining factor. Consider again the passage quoted earlier:

> Once C is recruited as a cause of M – and recruited as a cause of M because of what it indicates about F – C acquires, thereby, the function of indicating F. *Hence* C comes to *represent* F. *C acquires its semantics, a genuine meaning, at the very moment when a component of its natural meaning (the fact that it indicates F) acquires an explanatory relevance* (1988, 84, emphasis altered).

For Dretske, the "explanatory relevance" is, of course, the forging of the causal link between *C* and *M*. It is this which sees Dretske lose his grip on the causal efficacy of content, for it is only the recruitment of *C* to play a *role* in the organism's cognitive economy which *confers C's content*. Antecedent to its cognitive role, *C*, for Dretske, doesn't represent, or possess "genuine meaning". And, to put it mildly, it's not straightforwardly apparent how to assert an explanatory role for content if the content-determining factor is itself *secured* by the *effects* of representation bearers. Once he imports the causation of behaviour (*C*'s cognitive role) into his story about content determination, Dretske has stepped over into territory where content is no longer causally potent, and, as suggested, this constitutes a problem because it is in conflict with his own explicitly stated assumptions.[41]

---

[41] See Chapter Four, subsection 1.4. for a more detailed version of this kind of argument leveled against functionalist accounts of content determination.

The above difficulty has damaging consequences for Dretske's purported solution to the problem of misrepresentation. In essence, the previous criticism is that Dretske has not secured a causal role for *content* (functional meaning), merely a causal role for *indication* (natural meaning). That is, *C*'s causal role (the fact that it causes *M*) is determined by what it indicates, not by its content. I'm now going to claim that Dretske cannot secure *determinate* content, that his theory does not solve the disjunction problem.

In order to solve the disjunction problem, recall, what Dretske needs to show is that it is *C*'s indication of just *one* condition, namely *F*, which explains *C*'s causing *M*. What he needs to show, in other words, is why *C*'s content is "*F*" even though *C* covaries with other conditions (*G*, and *H*) as well as *F*. Dretske's proposal, again, is that *C* acquires its content only when it is recruited as of cause of *M* because of its indication of *F*. Now, as I argued above, Dretske has secured a causal role for indication, not content. But if it's (merely) indication that has a causal role (*C* causes *M*), what reason has Dretske given us for claiming that it is *C*'s indication of *F*, rather than *F-or-G-or-H*, which explains *C*'s causing *M*? And if it's *C*'s indication of *F-or-G-or-H* which explains *C*'s causing *M*, then because *C* only has *content* by virtue of *C* causing *M*, *C*'s content is not "*F*", as Dretske claims, but "*F-or-G-or-H*". Dretske has not solved the disjunction problem. His purported solution relies on assuming a privileged indication relation between *C* and *F*, something he is not entitled to do.

Dretske might respond to the above criticism in the following way. It is *C*'s indication of *F*, and not of *F-or-G-or-H*, which is privileged (and so "*F*" is *C*'s content), because *F* is what the organism wants or needs; *F* contributes to the organism's survival. *F* is, for example, a creature the organism feeds on, while *G* is an inedible creature similar in appearance to *F*, while *H* is merely a shape roughly similar in appearance to both *F* and *G*.

There are, however, two problems with such a reply. First, to use an expression Fodor (1990a, 67) has employed in a similar context, Dretske can only make this claim because he's "sneaked a look" at the environmental condition he needs in order to make his TCD work in the way he wants; in a manner, that is, which solves the disjunction problem. He has no principled reason for claiming that it is *C*'s indication of *F* which is privileged and which can be identified as *C*'s content.

Second, it may well be the case that the indication of *F-or-G-or-H*, rather than just *F*, better promotes the organism's survival. The reply that *F* is what the organism wants or needs – the idea that it is the indication of *F* which is survival-enhancing – is thereby undermined. Godfrey-Smith (1992) has argued that certain cost-benefit considerations will often weigh in favour of an organism employing a detector (*C*) which reliably covaries with, indicates, *F-or-G-or-H*, rather than just *F* (and thus the content of *C* is *"F-or-G-or-H"*, not *"F"*). If *M* is a relatively easy and unrisky behaviour, then, given what I've stipulated *F*, *G*, and *H* stand for, it's better for *C* to cause *M* (assume *M* is some sort of pursuit behaviour) even if it's *G* or *H* that turns out to be present, rather than having a detector which causes *M* only when *F* is present (or almost certainly present). For a detector to fire (and to cause *M*) only when *F* is present – to be more reliable – then presumably it would have to fire at a later time than an *F-or-G-or-H* detector. And if this is the case, creature *F* ("lunch") will have a greater chance of itself detecting the predator and escaping. So, contrary to Dretske's possible reply, there are reasons for thinking that *C*'s indication of *F* cannot be privileged on survival-enhancing grounds.

It is clear that the very character of Dretske's TCD has been strongly influenced by the need to incorporate the capacity for misrepresentation. However, though Dretske has developed a sophisticated proposal for overcoming the problem of misrepresentation, the result, as I've attempted to demonstrate, is not successful. I want to finish this subsection with a final general comment about Dretske's theory.

To my mind, Dretske's notion of functional meaning is somewhat overly "theoretical". The construct of functional meaning seems more like an artefact of – an *ad hoc* manoeuvre born out of – the desire to accommodate misrepresentation, rather than an independently plausible construal of cognitive content. There is more at work here, however, than a clash of intuitions. As touched on above, Dretske equivocates in regard to the bearers of natural (or biological) functions. He slides between attributing functions to systems, devices, or *mechanisms*, and attributing functions to the *states* of such systems.[42] However, as a number of writers have pointed out, there is a significant difference between attributing a biological function to a mechanism (which produces cognitive states), and attributing a

[42] In his 1995 work (for example, 2, 169 note 1) Dretske attributes indicator functions to both states and mechanisms.

biological function to particular states of this mechanism (see, for example, Godfrey-Smith 1989, 542; Fodor 1990a, 65-66). For one thing, while it's plausible to suppose that cognitive mechanisms have biological functions, it seems much less plausible to suggest that particular cognitive states have natural functions.

More importantly, as Dretske realises (referring to a point made by Godfrey-Smith 1989, 542), "it is systems (for producing states), not the states these systems produce, that are subject to the process of natural selection" (1995, 171 n. 11). This, however, creates a difficulty for Dretske's account of content determination. His account requires that cognitive-*state* types have biological functions, and that these particular functions are subject to the process of natural selection. If it is only cognitive mechanisms that have biological functions, then Dretske doesn't have a notion of function which cuts fine enough to deal with the grain of mental content. That is, mental content slices finer than the functions of cognitive mechanisms. A given cognitive mechanism has the capacity to go into, or produce, a whole host of cognitive states; states with different contents. The visual system, for example, we might be prepared to say, has the biological function of providing information about the environment carried by light. Such a system is capable of producing an enormous number of states with different contents. The etiology of this system, however, does not confer a biological function upon any one of these states. Only structural features of the visual system are products of an evolutionary history (Godfrey-Smith 1989, 542). In order for his account of mental content determination to work, Dretske would have to make the case for particular states, not merely cognitive mechanisms, possessing biological functions. However, it seems clear that appeals to natural selection won't help in this regard. And, as shown above, Dretske's attempt to secure determinate biological functions in terms of a representation bearer's characteristic behavioural effect is also beset by problems.

While Dretske has worked hard at the problem of misrepresentation, I remain unconvinced that his proposed solution succeeds. So we are still without a plausible means of dealing with the capacity for misrepresentation. Other philosophers, however, have thought that Dretske is on the right track with his appeal to teleological considerations. I turn to a TCD in which teleology figures centrally in the next subsection.

2.4. Millikan

While Dretske adds a teleological component to indication, a purely teleological TCD has been attracting a growing number of adherents (see, for example, Millikan 1984, 1993; Neander 1991a, 1991b, Papineau 1987, 1993). In order to establish the claim that there is a common general strategy adopted toward content determination and misrepresentation, it is therefore necessary to consider the approach often referred to "teleosemantics". I will focus on the work of Millikan, the most prominent defender of a teleological TCD. Since I've already administered a good dose of teleology in the discussion of Dretske's work, I will provide only a very rough sketch of Millikan's particular employment of teleological resources.

Millikan emphasises that mental representation is a biological phenomenon, and, in her view, content determination is a biological matter. She begins with the (uncontroversial) claim that cognitive systems are a product of evolutionary forces which generate and utilise representation bearers in order to interact appropriately with their environments (1993, 11). Content determination is a matter of biological function, where 'biological function' is understood not in terms of what a state (or process[43]) is currently doing, but in terms of what it is supposed to do. 'Supposed to do' is, in turn, unpacked by appeal to evolutionary "design": what a state did that accounted for its selection, and for the survival and proliferation of the organisms possessing the state type. Millikan employs the expression "Proper function" to refer to this historical rendering of 'function'. A state type of an organism has a Proper function if: its function is to indicate that a particular environmental condition obtains; such a state produces a characteristic behaviour adaptive to the presence of the particular condition; and such a survival-enhancing co-ordination of behaviour with environment was present in the organism's ancestors.

So, according to Millikan, biological systems possess states (mechanisms, and processes) having Proper functions. The Proper function of a state type is a capacity responsible for its continued existence. A state type's Proper function, that

---

[43] The parenthetic remark is important in Millikan's case because she has been the foremost advocate of the view that devices or processes which use, or "consume", representation bearers must be incorporated into a satisfactory analysis of content determination. Consequently, Millikan distinguishes between "representation producers" and "representation consumers" (see, for example, 1993, Chps. 4 and 6). Having noted this distinction, I will, for the sake of brevity, gloss over it.

is, is what it was selected for. It may not now be exercising, and have only rarely in the past exercised, such a capacity. Proper function, then, is a normative, rather than statistical, notion.

The notion of Proper function is the basis of Millikan's TCD.[44] The content of a state type, according to Millikan, is that environmental condition it is the type's Proper function to represent. When operating as it is supposed to – as it has been selected for – a biological mechanism produces a state which maps onto a certain environmental condition, and it is that environmental condition which is identified as the state's content.

For Millikan, misrepresentation occurs when a state token is produced by an environmental condition other than the condition the state *should* represent. If circumstances are other than those which, in the past, resulted in a certain state mapping onto a certain environmental condition, then typically the result is misrepresentation. Consider the much-discussed example of a frog snapping at a BB pellet. On Millikan's account, the content of the frog's representational state responsible for the feeding-directed response is "bug" (or "fly", or perhaps "food" or "prey"; though, obviously, not rendered in terms of an English language expression for the frog). The state which produces snapping behaviour is a misrepresentation because it is tokened by, and thereby applied to, a BB pellet, not a bug. According to Millikan, the content of the frog's internal state is determined by the state-type's Proper function. As suggested, the Proper function of a representational state is the representation of that condition which contributed to the survival and proliferation of the organism's ancestors. It was the state's representation of bugs, and the state's production of behaviour appropriate to the presence of bugs, which contributed to the adaptive success of the frog's ancestors. So, in this case, the content of the state is "bug", rather than "bug-or-BB-pellet" or "small-dark-moving-dot". The representation of bugs is what the state was selected

[44] There are other aspects of Millikan's TCD which are associated with the idea of Proper function, including the notions of "normal explanation" and "normal condition". On the former, Millikan writes: "A 'normal explanation' explains the performance of a particular function, telling how it was (typically) historically performed on those (perhaps rare) occasions when it was properly performed" (1993, 86). On the latter, Millikan says: "A 'normal condition for performance of a function' is a condition the presence of which must be mentioned in giving a full normal explanation for performance of that function. ...For example, normal conditions for discriminating colors are not the same as normal conditions for discriminating tastes....[N]ormal conditions might almost better be called 'historically optimal' conditions" (1993, 87). For more detail see Millikan (1984, Chps. 1, 2, and 6; 1993, Chp. 4).

for. Tokening the state with content "bug" in response to a BB pellet is, then, an instance of misrepresentation.

The general approach to content determination and misrepresentation illustrated above in the discussion of Fodor and Dretske is, as suggested, also adopted by Millikan. Consider the following passages from Millikan:

> If circumstances are not normal for the [representation-] producer, that is, if circumstances are not as they have historically been when the [representation-] producer has succeeded in its task, then,...the [representation-] producer will almost undoubtedly fail. For example, the frog's bug-detecting apparatus fails whenever a BB shadow crosses the frog's retina. It fails to produce a true representation (1993, 130).

And:

> ...inner representations are produced by systems having as teleofunctions the production of *true* representations, and...misrepresentation occurs when these systems miss (1993, 129, emphasis added).

Like Fodor and Dretske, a TCD, for Millikan, describes the circumstance in which the determination of veridical content of state types occurs. Misrepresentational tokens (which have their content by virtue of the circumstance conferring content to the type) occur in circumstances other than the content-conferring circumstance. Again, the distinction between type one and type two situations can be employed. Content determination occurs in the specified type one situation, and tokens so produced are necessarily veridical. For Millikan, a type one situation is a teleologically normal one. Tokenings in such a situation represent that environmental condition which, historically, it was adaptive for the organism's kind to represent. When conditions are other than teleologically normal – in the case, that is, of a type two situation – misrepresentation occurs.

Again, like the accounts of Fodor and Dretske, Millikan's proposal for misrepresentation is an augmentation of the basic TCD story; it resides within the account of content determination. Millikan uses the same theoretical tools to explicate both content determination and misrepresentation. Content determination, as shown above, is underwritten by appeal to Proper biological function. The content of a state type is that environmental condition it should represent, that condition it was "designed" (by natural selection) to represent. Misrepresentation is explained in terms of a token state failing to perform its Proper function. A token representation bearer misrepresents when it doesn't map onto the condition which, in the organism's ancestors, tokens of such a type were historically mapped onto such that a survival advantage resulted. On Millikan's theory, misrepresentations are, in some sense, maladaptive.

With that rather brief presentation of Millikan's theory, I now turn to some objections. The first turns on Millikan's idea that correct representation – veridical content determination – is to be unpacked in terms of adaptiveness. The problem for Millikan here is that correctness of content does not straightforwardly line up with adaptive outcomes. On the one hand, we have the sort of situation mentioned above in the discussion of the difficulties with Dretske's more recent theory: sometimes misrepresentation holds a survival-enhancing advantage over correct representation (Godfrey-Smith 1992; Cummins 1996, 45). The idea can be illustrated with the following examples. The gazelle that bolts at a rustle in the bushes caused by the wind may look the fool, but falsely representing such a condition as the presence of a predator is adaptive when enough rustlings (and it needn't be very many when the potential consequences are so dire) are due to the presence of a lioness. Or consider a scenario due to Cummins:

> [There are] cases in which...errors are not *serious*. A system that consistently represents little ambient black particles as insects will serve a trout well in an environment in which enough of the little ambient black particles are insects, and few are harmful. It is cheaper to design a tolerant digestive system in such circumstances than it is to design an insect recognition system free of false positives (1996, 45).

So, not only correct representations are adaptive, and this makes trouble for Millikan's attempt to reduce content determination to adaptiveness.

On the other hand, representational content can be veridical without being adaptive. Again I borrow an example from Cummins: "You can design a trout that, in spite of refraction, correctly represents the positions of insects flying just above the water, but this will not be adaptive in a trout already equipped with a jumping routine that compensates for refraction" (1996, 45). The examples here might seem unrealistic, as they deal with rather simple representational creatures. But Millikan herself draws heavily on primitive representational capacities found in very basic organisms, so I am not doing her theory any disservice here.[45]

The second main criticism of Millikan's proposal is due to Fodor (1990a), and was hinted at in the discussion of Dretske. The charge, in outline, is that natural selection is too blunt an instrument to provide for the determination of determinate content. There are two closely related claims in Fodor's (1990a, Chp. 3) first argument against the attempt to underwrite content determination in terms of Normal function, and, in turn, selectional history. The first concerns the indeterminacy of biological functions. The second focuses on the idea that appeals to the Proper function of a mechanism cannot distinguish between an organism's representation of a certain environmental object and an organism's representation of a different environmental object when the two objects are correlated in an organism's environment. The second claim is a consequence of the first.

A Proper function is, of course, specified in terms of a particular environmental condition. It's the Proper function of mechanism *M* to respond, for example, to the presence of flies.[46] The problem with specifying function in terms of the content or representational object of the state is, as Fodor points out, that function goes wobbly whenever there is uncertainty about the state's content (1990a, 70 ff). That is, there is an alternative construal of the function of the frog's internal mechanism because there exists an alternative way of rendering that which the mechanism represents. It's perfectly plausible to suggest that what *M* is selected for is the detection of not flies, but little ambient black things. The reason it's perfectly plausible is that it's perfectly consistent with teleological explanation to describe *M*'s content as little ambient black things when, in the environment in which *M* operates, enough of the little ambient black things *are* flies. Little black ambient things occasion representational tokens which induce snapping behaviour, and since froggie gets enough flies into him by snapping at little black ambient

---

[45] Indeed, one might conjecture that Millikan relies too heavily on very primitive representational capacities (those tokenings of representational states which occasion a "hard-wired' behavioural response), in that it is not explained how her theory is meant to "scale up" to more sophisticated representational creatures such as ourselves. That is, Millikan's theory looks more implausible the further up the ladder of evolutionary complexity we scale.

[46] Because, in the Normal circumstance, flies occasion "fly" representation bearers, and "fly" representation bearers bring about snap-at-fly behaviour; this is what *M* was selected for. So *M*, or anyway, a state therein, represents flies. Incidentally, nothing turns on the change in the example from the representation of bugs to that of flies. It is purely expository; Fodor's argument being related here employs the example of flies.

things – because enough of the little black things froggie snaps at are flies – *M* is going to be selected for its registration of little black ambient things. So, the content of the state that causes froggie to snap is "little black ambient things".

The interpretation Millikan wants is that *M* is functioning Normally when it produces snaps at flies, in which case snaps at BB pellets are errors. On the alternative interpretation, *M* is functioning Normally when it produces snaps at little black ambient things, in which case snaps at BB pellets are not errors. The problem for teleosemantics is that evolutionary theory provides no principled way of ascertaining which of these interpretations is privileged. Here's how Fodor puts it:

> The Moral…is that…Darwin doesn't care how you describe the…[representational] objects of fly snaps. All that matters for selection is how many flies the frog manages to ingest in consequence of its snapping, and this number comes out exactly the same whether one describes the function of the snap-guidance mechanism with respect to a world that is populated by flies that are, de facto, ambient black dots, or with respect to a world that is populated by ambient black dots that are, de facto, flies. … *Darwin cares how many flies you eat, but not what description you eat them under* (1990a, 72-73).

Since there is no fact of the matter regarding *M*'s function, its function is indeterminate, and thereby whether *M* represents flies or *M* represents little black ambient things is indeterminate. And if *M*'s content is indeterminate, then Millikan doesn't have a solution to the problem of misrepresentation.

The second claim, as suggested, is closely associated with, and follows from, the first. The problem just outlined is that Proper function is indeterminate since there's no principled way of deciding which environmental object *M*'s function incorporates. As a consequence, a specification of *M*'s function will be ambiguous between any two or more different environmental objects which are reliably correlated in froggie's environment (see Fodor 1990a, 73-74). That is, if *F* (flies) and *B* (BB pellets) are correlated in froggie's pond, the adaptiveness of representing *F* and of representing *B* turns out the same. In terms of selection, the representation of *B* does just as well as the representation of *F*. Evolutionary teleology cannot distinguish between reliably correlated contents, content is thereby indeterminate, and so the teleological approach to content determination and misrepresentation doesn't resolve the problem of representational error.

I want to close this subsection with a comment that goes to the general character of the TCD at hand. Teleological accounts of content determination look to get the *explanatory order* the wrong way around (Cummins 1996, 46). Such accounts seek to reduce content determination to Proper function, and Proper function to adaptational history. But surely representation bearers are selected *because* of what they *represent*. On the causal theory, a representation bearer is selected, one can say, because it represents an environmental object, and it represents that object by virtue of being lawfully correlated with that object. On a second-order resemblance TCD, a representation bearer is selected because it is part of a representing structure which is second-order isomorphic to a represented structure. The more fine-grained the isomorphism, and the more representing structures the organism has, the greater its adaptability. Representation bearers are selected because of the mapping, their mapping is not determined by their selection. It seems straightforwardly wrong to argue that what representation bearers represent is a matter of their adaptiveness, their selectional history.

The mistake is a product of beginning with the determination of veridical content, coupled with the plausible sounding assumption that representation bearers are typically (though, as indicated, not exclusively) adaptive when veridical. The reason why the assumption sounds plausible is that representation bearers are adaptive *because* they are veridical. And you can't make a case for this idea if you hold, as teleologists do, that what a representation bearer correctly represents is what it is adaptive to represent. Teleologists start with the determination of veridical content, and then have to find a principled means of saying how the content of a representation bearer token, and the object it is applied to, come apart. They propose to do so by saying, firstly, that a representation bearer's content is correct because it is applied to the object it is adaptive to represent, and secondly, a representation bearer's content is in error because it is not applied to that object it is adaptive to represent. But the right thing to say is that a representation bearer is adaptive, or selected, because it represents veridically, and not adaptive when its content is in error. You can't, however, say this if representational content is determined by what's adaptive. Content must be determined independently of what's adaptive so that representation bearers (or, more precisely, the systems which generate them) can be selected for their representational capabilities.

Adaptive success is accounted for in terms of representational success. Content determination is explanatorily prior to adaptiveness.

The above objections to teleosemantics and its proposed solution to the problem of misrepresentation demonstrate that Millikan does not currently have a convincing theory of mental content determination on her hands. Given the critical difficulties facing the purported solutions of Fodor and Dretske, the theorist looking to overcome the problem of misrepresentation is left in something of a bind: either revisit one or more of the above proposals and attempt some modification, or strike out in a new direction. To my mind, the former option does not look promising. I've sought to demonstrate that despite the variations among the standard proposals for content determination and misrepresentation, they exhibit a shared set of characteristics and commitments, which allows me to propose the following suggestion: it looks as if the very *character* of the orthodox approach is the *reason* why misrepresentation has proven particularly recalcitrant. Consequently, it would be useful to highlight that character in order to reveal the shape of an alternative approach. That is the task I have begun here. The next chapter embarks upon the construction of what I propose as a replacement strategy. I will further sharpen the picture of the proffered replacement by isolating additional characteristics of the established approach. That is where the next chapter begins.

CHAPTER THREE

INTENTIONALITY AND REPRESENTATIONAL ERROR

The current chapter continues the investigation into the phenomenon of mental representation begun above (Chp. One, sec. 2.), and further pursues the orthodox approach to content determination and misrepresentation. The focus will be on possible conceptions of intentionality, representational content, and content determination, as well as representational error. The literature on mental representation suggests that there is widespread agreement in regard to how intentionality and content should be understood. There has also emerged something of a consensus concerning the broad shape of an answer to the question of content determination (at least among those advocates of original intentionality). In section 1., I will suggest that current thinking about mental representation incorporates a particular understanding of intentionality, content, and content determination, and set about describing these further characteristics of the orthodox approach. Indeed, I intend to argue to the effect that the usual interpretation is in some part responsible for the orthodox methodology in regard to content determination and misrepresentation. I will propose an alternative way of understanding intentionality, content, and content determination, and subsequently show how such an alternative conception points toward a new framework for the accommodation of misrepresentation.

Section 2. provides a summary of the issues considered in Chapter Two and section 1. of the present chapter. I intend to rehearse the traits of the orthodox approach in light of the emerging alternative strategy. By the close of the current chapter, the character of the alternative framework will be fully revealed as the assumptions and commitments of the orthodox approach are relinquished and replaced.

1. Intentionality and Representational Error: Two Conceptions

In this section I will investigate two conceptions of intentionality, and, concomitantly, two contrasting views of representational content and content determination. One conception is the prevailing understanding; the other, I suggest, is a viable alternative notion (subsecs. 1.1. and 1.2.). I then show that the contrast between two perspectives on representational content does, in turn, admit of two possible interpretations, a *type* and a *token* interpretation. This distinction lines up with two kinds of representational error, respectively, selectivity and misrepresentation (subsec. 1.3.). I move on to examine a distinction crucial to the phenomenon of misrepresentation (first broached in Chp. Two, sec. 1. above), that between *content* and *application* (subsec. 1.4.). It is here that the heart of the alternative framework for accommodating misrepresentation starts to emerge. For the remainder of the section, I put forward certain considerations which, in my view, are responsible for the prevailing conception of intentionality (subsec. 1.5).

1.1. Intentionality: Two Conceptions

There is a consensus concerning the *meaning* of 'intentionality': intentionality is "aboutness" (Dennett 1987, 240; Fodor 1987, xi; Searle 1981, 1.). That is, 'intentionality" is uncontroversially taken to refer to the idea that the bearer of intentionality is about, stands for, refers to, points to, or is directed at, some other feature, object, or state of affairs. However, I suggest that the phenomenon of intentionality – the property of "aboutness" – can be understood in either of two ways, which I will label "intentionality-as-connectedness" and "intentionality-as-directedness". Each understanding of intentionality incorporates a particular understanding of the notions of content and content determination.

To conceive of intentionality as "*connectedness*" is to hold that mental representation bearers in some sense connect or make contact with objects in the external, mind-independent world. Intentionality consists of a real connection between a representation bearer and its representational object.[47] On this understanding of intentionality, the content of a representation bearer (its specific intentionality) is the object in the (represented) world, considered objectively, with which the bearer connects. And content determination is a matter of an *externally-instituted* objective relation obtaining between an object in the (represented) world and a representation bearer. To construe intentionality as "connectedness" is to

[47] Fodor sometimes employs the locutions "connect with" and "make contact with" in discussing the mind's relationship to the external world (see, for example, 1994, 83, 87). The thinking seems to be that *only* by internal states connecting with the external world can we explain behavioural success. I'll return to this point below (Chp. Five, sec. 3.).

incorporate the representational-object-*itself* within the phenomena of content and content determination.[48] The representational-object-itself is the represented object as it exists independently of its representation. The object, in other words, is considered independently of the *way* it is represented. It is to be distinguished from the representational-object-*as-represented*.[49] (To avoid overly unwieldy expressions, I will often employ the abbreviations RO-itself and RO-as-represented). I will return to this contrast below. On this view, then, intentionality is akin to reference, or denotation.

Intentionality-as-connectedness, I submit, is currently the orthodox view, found in the influential theories of Dretske, Fodor, and Millikan. According to Fodor's causal TCD, the content of a representation bearer type is the property – considered objectively, or in and of itself – which, if instantiated, produces a bearer token of that type. The content of a representation bearer is what it denotes. And content determination is externalist or wide, explained in terms of counterfactual causal relations obtaining between property instantiations in the external world and representation bearer tokenings. Millikan and Dretske likewise identify content with the RO-itself. In general terms, the content of a representation bearer type is the environmental object the bearer type has the function of indicating. And content determination is again wide, understood as a teleological relation between an environmental object and a representation bearer type.

An alternative way of conceiving of intentionality is in terms of "directedness". This is, in a sense, a weaker notion than "connectedness". To conceive of intentionality as directedness is to hold not that representation bearers connect or make contact with environmental objects, but merely that they are directed at, or point to, certain environmental objects. Or, more precisely, representation bearers point to *purported* environmental objects; they need not possess objective referents. On this understanding, it is the mere act of pointing beyond itself to a purported environmental object which constitutes the property of intentionality. In contrast to the understanding of intentionality as connectedness,

---

[48] I owe the term "representational-object-itself" to Von Eckardt (1993).

[49] I am not suggesting that the "connectedness" conception is committed to understanding the object represented as the representational-object-itself out of necessity. Rather, understanding what is represented as the representational-object-itself is a *characteristic* of the "connectedness" conception (and thereby the orthodox approach). It may be possible for an advocate of the orthodox approach to construct a notion of the representational-object-as-represented which is compatible with the externalist account of content determination. However, I will subsequently suggest that such a task looks to be rather problematic.

the content of a representation bearer is not the representational-object-itself. Rather, on the directedness conception, content is understood in terms of the representational object *as it is represented* by the representation bearer. At a first pass, content consists of selected features purportedly possessed by purported objects. Representational content is the construction of a purported, or notional (Dennett 1987, Chp. 5), world. Content is "relation-*like*". And content determination does not consist of an externally-instituted relation holding between a representation bearer and an environmental object. Rather, the content-determining relation is instituted by the brain, by the representation bearers themselves. Content determination is internalist, or narrow. Intentionality, understood in this way, is akin to sense.[50]

I now wish to say a little more on the content and content determination components of the two conceptions of intentionality. I'll begin with content.

1.2. Content: Two Conceptions

Although, as suggested, a particular conception of representational content has recently come to prevail, the claim that there are two ways of understanding the phenomenon does sometimes surface in writings on mental representation[51]. Some time ago, Pierce (1931-58, 4.536) distinguished between the "Immediate Object" and the "Dynamical Object." Previously (Files 1996, 410 n. 10), I have expressed

[50] I have a number of comments concerning the terminology employed in this paragraph. First, I will have more to say on the idea of a "notional world" in Chapter Five, where further development and defence of the "directedness" conception takes place. Second, a "relation-like" understanding of intentionality was, interestingly, was held by Brentano (1974/1973). For Brentano, such an understanding was motivated by the fact that our intentional states can be about nonexistent objects. And third, I use the terms "sense" and "reference" in discussing the two conceptions of intentionality with some trepidation. They have a contentious history in the philosophy of language. Certain philosophers of mind have appropriated the terms, and there seems to be a rough intuitive feel for their meanings. Fodor, for example, often describes his account of content determination as "referential" or "denotational" semantics.

[51] Sometimes, as with so called "two-factor" TsCD, the claim is that there are two *components* of content. The "narrow" component is internally determined (typically by the internal causal roles of representation bearers), and is standardly thought to supervene on the brain. Because it is internally determined, and taken to correspond to the *way* the object is represented, it can be likened somewhat to what I have termed the "RO-as-represented".(There is, however, a significant difference between the standard interpretation of the "narrow" component of content and my rendering of the notion, as will become apparent below). The "wide" component of two-factor TsCD is standardly thought to be determined by a world-head causal relation. It does not supervene on the brain, and is understood in terms of the referent or truth condition of a mental representation bearer. It is something like what I have termed the "RO-itself". For two-factor theories, see Block (1986), Field (1977, 1978), Loar (1981, 1982), and McGinn (1982).

this contrast as above, that is, in terms of the distinction between the "representational object *as represented* by a representation bearer" (RO-as-represented) and the "representational object *itself*" (RO-itself). This is the terminology I will typically employ. In a like vein, Lloyd (1989, 14) has distinguished a representation bearer's "explicit content" from its "extensional content". The former is the "information the representation [bearer] itself conveys, independent of any fact about the real object or situation represented (if any)." The latter is the "actual objects(s) or event(s), under any description, which the representation [bearer] denotes." Dretske (1995, 23) utilises "sense"/"reference" terminology to express the distinction, which he sees as one between represented properties and represented objects: "Representation [bearers] have a sense (the properties they have the function of indicating) and, often enough, a reference (an object whose properties they represent)...." He illustrates the distinction with an example due to Nelson Goodman:

> The difference between represented object and represented property, between the reference and sense of a representation [bearer], is the same distinction Nelson Goodman (1976) was getting at in his contrast between a picture of a black horse ("black horse" here specifying the object the picture is a picture of) and a black-horse picture ("black-horse" here specifying what the picture depicts the object as) (Dretske 1995, 23-4).

The represented property/represented object contrast, I think, is a useful *initial* way of illustrating the RO-as-represented/RO-itself distinction. The idea that representation bearers capture certain properties or features of objects and not others – the idea that representation bearers or representing systems are *selective* – is a fundamental characteristic of representation in general. It pertains to the variable accuracy of different types of representation bearers, and different kinds of representing systems. Representation bearers don't represent objects *tout court*. They represent objects as being a certain way, as possessing certain features. Consider a representation of you by means of a wax statue (like those found in Madam Tussard's museum), a colour photograph, a black and white photograph, and a very young child's monochrome drawing. Obviously, the different kinds of representation bearers here are able to express different features because of the different representational media employed. Equally obviously, there is a progression from a relatively expressive kind of representation bearer to a rather impoverished one (assuming that our young artist is not extremely precocious).

Minds, I think it's clear, also exhibit such specificity to certain features of the representable world. Minds, or more particularly, complex mental states – *coalitions* of mental representation bearers – represent objects (with 'objects' understood literally). And it must be acknowledged, I believe, that in so doing, such complex mental phenomena represent certain features of objects while being silent on others. Here is a clear case of the distinction between the RO-itself (the object as it exists objectively, with all of its properties), and the RO-as-represented (the object as it is represented by representation bearers, the properties the object is represented as exhibiting).

However, I don't think that the object/property distinction, *at bottom*, lines-up with the difference between the RO-itself and the RO-as-represented. The reason, in overview, is that I think we need to move to a conception of what is represented more fine-grained than that of whole objects. There is good reason to suppose (as I'll show in Chp. Four, sec. 2. below) that (simple or primitive) mental representation bearers are dedicated to the representation of particular features of objects and states of affairs. And to put a remark of Fodor's in generic terms, it is the primitive representation bearers of cognition that are at the bottom of the pile when it comes to the question of content determination (1987, 98). The point for now is that even at this more finely individuated level, there is a distinction to be drawn between the represented *feature itself* and the represented feature *as represented.* I'm not suggesting that the represented object/represented property contrast is useless or incoherent. As indicated, it is a viable means of first illustrating the RO-itself/RO-as-represented distinction (particularly at the level of complex representing items), and I rely on it further below. I'm claiming, however, that the distinction between two ways of understanding the representational object (remember I've explicitly signalled that I'll be employing representational "object" to cover states of affairs, object, feature, or value of a variable) does not *ultimately* bottom-out in object/property terms. I'll briefly return to this point in a moment.

The idea of the RO-*as-represented* sometimes arises independently of an explicit contrast with the RO-itself. Searle is perhaps the most ardent advocate, holding that *all* intentional states possess "aspectual shape" (1983, 13; 1992, 131, 155). The claim is that mental states represent things as being a certain way. Mental representation bearers capture, or present, certain aspects of objects and not others. Dretske (1995, 30-32) borrows Searle's "aspectual shape" terminology, and Lloyd (1989, 13-14) writes of the "specificity" of mental representation bearers. McGinn (1989, 190-192) talks of "modes of presentation". Likewise, Cummins (1996, 109) points out that representation bearers "capture certain [features] and are silent about others."

The idea of the RO-as-represented is perhaps clearest in the case of phenomenal consciousness, which is the domain Searle employs as an illustration:

> My conscious experiences, unlike the objects of the experiences [the representational objects themselves], are always perspectival. They are always from a point of view. But the objects themselves have no point of view. Perspective and point of view are most obvious for vision, but of course they are features of our other sensory experiences as well. If I touch the table, I experience it only under certain aspects and from a certain spatial location. If I hear a sound, I hear it only from a certain direction and hear certain aspects of it. And so on.
>
> Noticing the perspectival character of conscious experience is a good way to remind ourselves that *all intentionality is aspectual*. Seeing an object from a point of view, for example, is seeing it under certain aspects and not others. In this sense, all seeing is "seeing as". And what goes for seeing goes for all forms of intentionality, conscious and unconscious. All representations represent their objects, or other conditions of satisfaction, under aspects. Every intentional state has what I call *an aspectual shape* (1992, 131).

So, the fact that sensory experiences always occur from a particular point of view, a certain perspective, is one element of the idea of the RO-as-represented. Just by changing our position relative to an environmental object changes the content of, for example, a visual experience of that object. Another thing meant by the idea of the RO-as-represented, as already indicated, and as the above quotation makes clear, is that sensory experiences represent certain aspects of their representational objects. They represent certain features of an object, while failing to provide a representation of other properties the object may possess. Visual experiences, for example, don't represent an object's microstructure or its mass. A visual experience of a body of water represents it in a certain position relative to the perceiver, and as possessing certain shape, reflectance, textual, and perhaps movement properties, not as a collection of H_2O molecules.

Employing conscious states to illustrate the idea of the RO-as-represented, while useful, also has the potential to mislead. The RO-as-represented does *not* correspond to the qualitative character of experiences. An experience implicates a mental representation bearer; the qualitative character of an experience is a mental representation bearer under a phenomenological (as opposed to neurophysiological or computational) description. The RO-as-represented is what is represented of the object, those features or dimensions of the object which are represented by the isolable dimensions of our phenomenal experiences. Within some sensory modalities, the dimensions of phenomenal experience are known. Colour experience, for example, admits of three distinct dimensions, namely, hue, saturation, and brightness (with three further dimensions to specify colour at a certain location). Auditory experience has dimensions of pitch and loudness (again with three additional dimensions of location). The RO-as-represented consists in those properties of the external item (for example, the frequency of sound waves, the intensity of electromagnetic radiation) which are represented by the distinguishable dimensions of our phenomenal experience in the various modalities.

Searle, in my view, is right to stress the aspectual nature of our representational states. It's not only sensory experiences that exhibit aspectual shape. "Higher-order" mental states, thoughts, also represent their objects in particular ways. To use the time-honoured examples, we can think of Venus under the description "the morning star", or under the description "the evening star". We can think of a novelist in terms of his birth name, or his *nom de plume*. If I believe that it is raining (to borrow one of Searle's 1983, 13 examples), then I represent the current state of affairs as that "it is raining", and not, for example, that "the ground is wet", or that "water is falling from the sky". In general, mental states – mental contents – *slice finer* than the RO-itself.

The idea of the RO-as-represented, then, is undoubtedly important. Indeed, it is sometimes given the status of a *desideratum* any TCD must satisfy. (Cummins 1996, Dretske 1995, and Lloyd 1989 appear to see it this way, and McGinn 1989 clearly does). The crucial point is that mental representation bearers have content in this way (or this component of content), even if the object the representation bearer represents does not exist, or does not exist in the way the representation bearer represents it as being. Despite its importance, however, existing TsCD do not appear to be designed with the idea of the RO-as-represented in mind. It's one thing to have a TCD which, if pushed, may look capable of securing aspectual content. It's another to have a TCD which builds it in from the start. The TCD being developed in this work is explicitly aligned with a RO-as-represented understanding of content (as will become more apparent in Chapters Four and Five). Those TsCD which operate within the orthodox approach, I think, would experience particular difficulty in attempting to operate with a RO-as-represented conception of content. To show why, I must briefly return to the issue of why the RO-as-represented/RO-itself distinction ultimately is not to be unpacked in terms of the represented property/represented object contrast.

I pointed out above that Dretske (1995, 23-24) proposes to draw the RO-as-represented/RO-itself distinction in terms of the difference between represented property and represented object (with 'object' here taken literally). However, I suggested that while the property/object contrast is a useful first way of thinking about the RO-as-represented/RO-itself distinction,[52] the latter doesn't ultimately reduce to the former. The reason, as indicated above, is that there is good evidence for maintaining that individual mental representation bearers are dedicated to the representation of specific features or properties of the world, and at this more finely-individuated level there is a distinction to be drawn between the *way* a feature is represented and the feature *itself*. On the conception of mental content I begin to develop in Chapter Four below, our mental representational capacities are subserved by *systems* of representation bearers, with any one such system dedicated to the representation of a particular feature or variable of the world. Individual representation bearers, on this view, represent the "value of a variable". The RO-as-represented then becomes the "value" a representation bearer represents a variable as exhibiting (or currently exhibiting). The RO-itself is the "value" the variable itself exhibits. Moreover, a system of representation bearers is going to represent a variable at a certain grain, or resolution. So a determinant of the RO-as-represented is the grain or value-structure of a representing system, which represents the represented-variable-itself as possessing a certain value-structure. The RO-as-

[52] Moreover, as indicated, I do think that the property/object contrast is particularly relevant when considering complex mental states. Coalitions of representation bearers yield representations of whole objects, and different coalitions of representation bearers represent different subsets of properties possessed by the objects themselves.

represented, in contrast to the RO-itself, is not independent of the representing relation. And, in my view, a TCD which accords a crucial content-determining role to mental representation bearers themselves, as a resemblance theory does, is much better placed to account for the idea that mental states represent the world as being a certain way.[53]

Those TsCD which operate within the orthodox approach, in contrast, will be hard pressed to explain mental content understood in terms of the RO-as-represented. The orthodox TsCD, as shown in Chapter Two, suffer rather badly from indeterminacy problems. Whether content determination is explicated in terms of causal covariation, or in a teleological fashion, the orthodox theories have trouble securing the representation of a determinate type of representational object over a (possibly open-ended) disjunction of representational objects. Indeed, the indeterminacy difficulties confronting the orthodox TsCD are even more pernicious than the discussion of Chapter Two suggested. I considered only one general kind of indeterminacy, that which Godfrey-Smith (1989, 535-536) refers to as "vertical" indeterminacy. This is the kind of indeterminacy found with the disjunction problem, the problem of accommodating misrepresentation. Here the theorist seeks to privilege one object as the content-determining object from all the objects located at the same "distance" from the cognitive system, all those objects capable of bearing a causal or teleological relation to the cognitive system's internal states. There is also the difficulty, again borrowing Godfrey-Smith's term, of "horizontal" indeterminacy. In this case the task is to specify how far back along the content-determining relation from external world to mental representation bearer we are to locate the representational object. Consider the causal chain from the presence of fire to a FIRE representation bearer. Why is the representational object fire, "rather than the structure of ambient light, or the presence of fuel at ignition temperature, or patterns of excitation on the retina?" (Godfrey-Smith 1989, 536). Now, my reason for thinking that the orthodox TsCD confront particular difficulties with the RO-as-represented conception of content is easily stated. The orthodox TsCD suffer from indeterminacy problems with mental content understood at a rather *coarse* level of analysis (recall Fodor's example of the representation of the object-type

[53] There are *two* ways "representation bearers themselves" can determine their representational content, as I'll show below (Chp. Four, sec. 1.) in the detailed examination of second-order isomorphism. One is by virtue of their internal causal roles. The other is by virtue of their physical properties.

horse, and the much discussed case of the frog who snaps at flies and BB pellets). Indeterminacy worries create enough of a headache for the orthodox TsCD with a rather coarse-grained conception of mental content. Such indeterminacy concerns will, I think it's clear, become chronic as we move to a more fine-grained understanding of what it is that individual types of mental representation bearers represent. And, of course, the RO-as-represented conception is a more fine-grained conception than the prevailing RO-itself understanding.[54]

1.3. Representational Error: Misrepresentation and Selectivity

So far in this section, an important concern has been to distinguish between the RO-as-represented and the RO-itself. I've pointed out that the notion of representational content (the object represented, or the representational object) is ambiguous between the object as it is represented as being, and the object as it is objectively. I've suggested that it's possible to identify representational content with the RO-as-represented, in contrast to the prevailing understanding which identifies representational content with the RO-itself.

What has gone largely unrecognised, however, is that there is another crucial distinction to be drawn. That is, expressions such as "what is represented", "the object represented", and "the representational object" are ambiguous in a further way. The ambiguity here is between "the object represented" understood as the *content* of a representation bearer, and understood as that to which a representation bearer is *applied*. So, we have not one, but two, distinctions. First, the content of a representation bearer can be understood as either the RO-as-represented or the RO-itself. If – as I'm proposing here – content is identified with the RO-as-represented, content is to be *distinguished* from the RO-itself. Second, the content of a representation bearer (whether interpreted as the RO-as-represented *or* the RO-itself) is to be distinguished from the object to which a representation bearer is applied. Importantly, the two distinctions concern two fundamental

[54] It might be maintained that the "language of thought" (Fodor 1975), coupled to one of the orthodox TsCD, could handle the RO-as-represented understanding of content for the propositional contents of "higher-order" mental states, certain beliefs and desires, for example. Primitive symbols can be combined in various manners to yield complex symbols which represent an object as being a certain way. That may well be so. But I find it difficult to see how the orthodox TsCD are to explain the aspectual character of our sensory and perceptual representational capacities.

characteristics of mental representation: *selectivity* and *misrepresentation*.[55] Over the next two subsections, I propose to develop the latter distinction, and clarify how it differs from the former.

The first distinction has been considered above. A brief recapitulation will aid the discussion to follow. In this case, the contrast is between the representation of an object as being a certain way, and the object itself which is being represented. That is, the RO-as-represented is the representation or expression by the representation bearer of certain aspects of an object. The RO-itself is the object purportedly represented, conceived in and of itself; conceived, that is, independently of its representation. The RO-itself is the actual object as it exists in the world, with all of its objective features. It is a token of a certain object-type. Consider a certain type of visual experience, for example, a horse-experience. The RO-as-represented of this experience consists in a certain subset of features – certain shape, size, texture, etc., features – possessed by the object-type horse. (It is very difficult to precisely describe the features represented by states of phenomenal consciousness. And describing them more fully in the terms of a natural language should not be taken to suggest that they are represented, in the mind, in a language-like manner). Such a subset of features "constructs" a purported object, a horse, at a certain location. The mental representation bearer "says", in effect, "There is, in such and such a location, (purportedly) an object with such a subset of features." The RO-itself is that object considered independently of its representation; an instance of the object-type horse.

The RO-as-represented/RO-itself distinction pertains to the *selectivity* of representation. Selectivity is, I suppose, representational error of sorts. It is a characteristic of mental representation related to, but not the same as, misrepresentation. With selectivity we are concerned with which features of the object-itself are picked-out by types of representation bearers, and whether these features are (in the object-itself) as the content of the representation-bearer type

[55] The latter distinction has not gone unnoticed by Cummins (1996). Others, however, appear to conflate the two distinctions, and thereby, selectivity and misrepresentation. Dretske (1995, Chp. 1.), for example, seems to see the RO-as-represented/RO-itself distinction (the "sense"/"reference" distinction in the terminology he employs) as relevant to the problem of misrepresentation.

"says".[56] I think it's time to recognise that a viable TCD must take seriously the aspectual character, or selectivity, of mental representation; this is especially so for a theory aimed at the contents of perceptual experience. And, as I pointed out above, it is not clear how the intentionality-as-connectedness approach can capture the selective character of mental representation. The fact is, again as suggested above, mental states *slice things finer* than the notion of the RO-itself. The intentionality-as-directedness approach takes this observation seriously, identifying the content of a representation bearer with the RO-as-represented. The RO-as-represented conception of content makes for a robust notion of the selectivity of representation. It also reveals that the RO-as-represented/RO-itself distinction – and, concomitantly, the selectivity of representation – concerns *types* (as opposed to tokens).

If content is identified with that subset of features an object is represented as having, representation bearer tokens belong to the same content-type in terms of representing the same subset of features possessed by a type of RO-itself.[57] Representation bearers are type-identified (by content, of course) by virtue of what subset of features they select. The content of a representation-bearer type is identified with the RO-as-represented *qua* type. Type-individuating by content, and against the connectedness (RO-itself) understanding, I can say that a type of representation bearer slices finer than a type of RO-itself. That is, any type of RO-itself can be represented in different ways. Different features of the RO-itself *qua* type can be captured, or represented, by different types of representation bearers. Or, different types of representation bearers select different subsets of features possessed by the type of object-itself. So, there is a many-to-one relation between (types of) mental representation bearers and (a type of) RO-itself.

Selectivity is bound up with an assessment of what I call the "accuracy" of representation. Representation-bearer types are more or less accurate depending on the number of features of the RO-itself (*qua* type) they capture, and at what degree of resolution. Selectivity, that is, incorporates the idea of the *variable accuracy* of

---

[56] Selectivity looks to be somewhat akin to the kind of representational error Cummins (1996, 23-25) terms "forced error". Error is "forced" when the representing system employed lacks the expressive power to adequately represent all aspects or dimensions of a certain representational domain. (Visual experiences, for example, are unable to represent water as H_2O molecules and twin-water as XYZ molecules). Forced error, of course, comes in degrees. Some representing systems have more expressive power than others in regard to certain representational domains.

[57] Or type*s* of representational objects themselves. A certain subset of features may be common to different types of objects.

different types of representation bearers. Assessing the accuracy of representation involves a comparison between representation bearers (type-identified in terms of various subsets of features they represent an object as having, at a certain resolution) and representational objects with their full complement of features.

I turn now to the second distinction, the second ambiguity in such expressions as "what is represented", "the object represented", and "the representational object". This distinction concerns *tokens*. Here, as foreshadowed above, the ambiguity is between the content of a representation bearer token and the token object to which the representation bearer is *applied*. On the latter interpretation, "the object represented" is the (token) object the (token) representation bearer is *used to represent*. It is what Cummins (1996, 6-8) calls the "target" of the representation bearer token. And as he writes, "...it is [representation bearer] tokens, not types, that have targets" (1996, 10 n. 3). The important point is that representation bearer tokens of the same content-type can be applied to token objects belonging to different object-types. A "horse" token, for example, can be applied to a cow, a donkey, or a horse-shaped bush, among other things.[58] A cow might cause the tokening of a "horse" representation bearer, and a cognitive agent, being cow confronted, thereby uses the "horse" token to represent the object confronting her (namely, a cow). So, we need to distinguish "the object represented" understood as the content of a representation bearer from "the object represented" interpreted as the object to which the representation bearer is applied. The distinction here, of course, pertains to the phenomenon of *misrepresentation*. We have misrepresentation when the *content* of a representation bearer does not match the object to which the representation bearer is *applied*.

1.4. Misrepresentation: Content and Application

In the previous subsection I suggested that the notion of the object represented – the representational object – is *doubly* ambiguous. On the one hand, what is represented

[58] Despite moving us in the direction of a more fine-grained conception of mental content than is usual, I've employed "horse" rather than "horse-aspects" as the content in order to emphasise that the distinction here is between the content of a representation bearer and the item to which it is applied. Such a distinction is legitimate whether content is identified with the RO-itself or the RO-as-represented. That is, someone might accept the importance of the content/application distinction to the problem of misrepresentation while disagreeing with the RO-as-represented conception of content being proposed in this work.

by a representation bearer can be understood as either the RO-as-represented or the RO-itself. The prevailing approach typically fails to recognise this ambiguity, and operates with the latter interpretation. The object represented is what a representation bearer refers to, or denotes. I've claimed, however, that a viable alternative notion of content is to be had in the form of the RO-as-represented. Indeed, the selectivity of mental representation suggests that the RO-as-represented conception of content is to be preferred. So content is to be distinguished from the RO-itself.

On the other hand, what is represented by a representation bearer is ambiguous between its content and the object to which it is applied. Again, the prevailing approach runs these two separate ideas together. On the orthodox conception, the object represented is that item a representation bearer is *used* to represent; content is conflated with (correct) application. It is natural to think in this way in the case of perception, where the environmental object that occasions a representation bearer tokening is the object the bearer is used to represent. It is but a small step to then identify content with the object a representation bearer is used to represent. Content and application are thereby run together. However, over the current subsection and the next, I aim to show that accommodating misrepresentation requires content and application to be kept distinct. I will also isolate some considerations which I think are in some part responsible for the conflation at hand, and the orthodox approach in general.

Cummins (1996) is one philosopher who has done much to highlight the latter of the two above-mentioned ambiguities. It is worth quoting him on this point:

> ...expressions like "what I referred to", "what I meant", and the like [for example, "what is represented", "the object represented", "the representational object"] are ambiguous. Sometimes they mean targets [intentional objects], sometimes contents. You say, "I used M [a map] and got around the city with no problem." "Which city do you mean?" ("Which city are you referring to?"), I ask. Here, I am asking for your *target*, the city against which M's accuracy is to be measured. "Here is the map of the city I was telling you about", you say. "Which city do you mean?", I ask. Here, I am asking for the *content*, that is, for the city the map actually represents (1996, 127, emphasis added).

In this passage Cummins distinguishes two senses of expressions such as "what is represented", "the representational object", and "referent". In one sense, such expressions are synonymous with representational content.[59] In another sense, they mean the target or intentional object. The "city against which M's accuracy is to be measured" (or better, the city against which the accuracy of the *use* of M is to be compared), is the city the map is applied to., the city the map is used to represent.[60] As suggested, it is this distinction between content and application which pertains to the issue of misrepresentation. Assessing the veridicality of representation involves comparing the content of a (token) representation bearer with the (token) object to which the bearer is applied. In the example, given the behavioural success – the successful navigation of the city – it seems the map was correctly applied. The map was applied to the city the map actually represents. Content matched intentional object, so we have veridical representation.

The crucial point is that the ambiguity in the above-cited expressions leads to a conflation of a representation bearer's intentional object and its content. Failing to distinguish the two creates a problem in accommodating misrepresentation. Any representation bearer token can be applied to different (intentional) objects, including objects it is not true of. To accommodate misrepresentation, the content of representation bearer tokens must be fixed *independently* of what they are applied to, so that tokens are thereby free to misrepresent when the intentional

---

[59] The contrast for Cummins (1996) is between the content of a representation bearer and its target (what I have called the intentional object). In regard to this contrast (see the previous note), the content of a representation bearer can be understood as either the RO-as-represented or the RO-itself. Apart from a few words pointing to what I've called the "selectivity" of representation, Cummins does not discuss the RO-as-represented/RO-itself distinction. In the above passage, "...the content, that is,...*the city* the map actually represents..." certainly seems to imply that content is being identified with the city *tout court*, and thereby the RO-itself.
Notice, however, that the map is also a nice illustration of the RO-as-represented/RO-itself distinction, and thus the *selectivity* of representation. The map does not represent all the features of the city, only some of them (the spatial arrangement of streets, certain buildings, for example). A topographical map selects a different feature of the city, as does one which shows population density by suburb. And different maps of the one kind can represent their aspects at various degrees of resolution.

[60] I'm providing an exposition of Cummins' point here, so I'm constrained by his terminology. Typically, however, I reserve the term 'accuracy' for discussion of the selectivity of representation (employing 'veridicality' when considering the issue of misrepresentation). Assessing the accuracy of a representation bearer (a coalition of representation bearers, or a representing system or systems) involves comparing the RO-as-represented *qua* type (content) with the RO-itself *qua* type. In the case of mental representation, because there is good reason to suppose that individual types of representation bearers are dedicated to the representation of specific features, an assessment of accuracy will often involve appeal to a number of representation bearers which collectively represent a whole object. That is, I take the accuracy of mental representation to refer to the issue of how many features of the object-itself are represented, and at what grain or resolution.

object (of a token application) does not match the content (of a given representation bearer tokening). What is required, then, is that the content of a representation bearer, and its intentional object, be determined independently. It is this claim which is at the heart of the alternative approach to content determination and misrepresentation.

The orthodox approach to content determination and misrepresentation fails to distinguish between content and application. The approach sees a representation bearer's intentional object as its content. And this means it fails to distinguish the determination of content from the determination of intentional object. Such an approach can, I think, be traced to a focus on the tokening relation in perception. The prevailing approach in constructing a TCD is to begin with the realm of perception, and view perceptual representation bearer tokenings as indicators (or detectors) of external objects.[61] It is in the realm of perception that it seems most plausible to think that what a representation bearer is applied to is the object before the cognitive system which brings about the representation bearer tokening. However, it is also natural to think, as suggested above, that what a representation bearer represents is (in one sense) the object confronting the cognitive system which tokens the representation bearer. This is certainly the understanding of the orthodox approach. A representation bearer in a detector system "fires" in the presence of an environmental object; it indicates – represents – that environmental object. However, if what a representation bearer is *applied* to is the object which brings about the representation bearer tokening, and what a representation bearer *represents* is also what causes its tokening, then what determines application and what determines content are the same thing. But, as suggested, application and content can not be run together, otherwise misrepresentation is problematic.

It is useful to ask why the prevailing focus is on detection; this will provide insight into why the orthodox approach is intentionality-as-connectedness with its problems in accommodating misrepresentation. That the orthodox approach conceives of content as the RO-itself and content determination as wide (externalist) can be partly attributed to operating (either explicitly or implicitly)

[61] I certainly don't think beginning with perception is a mistake. The problems for the orthodox approach begin after this point with the identification of content with the tokening object in cases of veridical perception.

with a "language of thought" conception of the bearers of intentionality.[62] On this conception, of course, bearers are symbols.[63] And symbols do not represent what they represent by virtue of their intrinsic properties. The relationship between the intrinsic properties of symbols and their representational contents is an arbitrary one. Symbols' intrinsic physical properties do not carry any information about objects they represent. It seems, then, that a LOT-style theorist has no option but to see perceptual representation bearers as detectors. And, as suggested, detectors simply "fire" in the presence of a certain environmental condition. It is only by virtue of being (reliably) tokened by a certain environmental condition that we can say that the content of the tokened representation bearer is that environmental condition. So, content determination is wide (externalist); content is determined by the environmental condition which causes the representation bearer tokening.

Operating with a LOT-style conception of representation bearers, then, results in a focus on the tokening relation and an understanding of (perceptual) representation bearers as detectors. As suggested above, in the realm of perception it is natural to think that the object a representation bearer is applied to is the object which causes the representation bearer to be tokened. Applying a representation bearer is using a representation bearer to represent something in the external world; application is relational. But the orthodox approach is to see the object which causes the tokening as the representation bearer's content. The content of a representation bearer is the object to which it is applied, on occasions of correct application (see Chapter Two). Application and content are thereby run together. And since application is wide, so too is content determination. However, it has been suggested above that collapsing the distinction between application and content determination leads to the problems with misrepresentation. What is required is that a representation bearer has its content determined independently of its (correct) application. If so, then it can be applied to token objects belonging to different object-types, including those which its content does not match. In light of the above

---

[62] I say "partly" because undoubtedly other factors have contributed to this approach. Foremost, perhaps, would be the standard intuitions about the morals of the Twin-Earth story (see, however, Cummins 1996, 123-128 for a nice rebuke to the standard intuitions). The recent prominence of the causal theory of reference in the philosophy of language might also be a factor. In attempting to construct TsCD for mental content, philosophers are often drawn to representation of other kinds in search of ideas.

[63] Fodor, of course, certainly constructs his TCD with a language of thought understanding of mental representation bearers. Dretske and Millikan, on the other hand, say little on the nature of representation bearers.

considerations showing the problem in seeing the tokening object as the determinant of content – if content is to be determined by something *other* than the (veridical) tokening relation – it seems there is only one way to go: content determination must be internal; that is, content must be determined by the brain's representation bearers themselves.[64] This, I suggest, is the only way we can presently envisage content determination in a way other than under the condition of correct application.

1.5. Connectedness, Directedness, and Veridicality

Why then, apart from the dominance of the view that mental representation bearers are symbols, has an internalist TCD (of the structural, rather than functional, sort) not been contemplated? To end this section, I wish to offer a few remarks which I think are part of an answer to this question. The answer turns on a distinction between a theory of *content* (TC) and a theory of *content determination* (TCD).

The *description* of representational content is what I will call a "theory of *content*". A theory of content describes, or explicates, what is represented. There are important issues concerning what representational contents the brain does encode, and how we go about stating the contents of certain sorts of intentional states, such as visual perceptions. My concern here, however, is that we do (or that it is in principle possible to) describe the contents of representational states. We can talk of that which constitutes the content, even if it is difficult to precisely express via language. A "theory of *content determination*", on the other hand, *explains* how content is conferred, or determined. The difference, then, between a TC and a TCD is between a theory which describes the relational (or relation-like) nature of content, and a theory which explains the nature of the content-determining relationship between representation bearer and representational object. The former theory is nonreductionist, the latter reductionist, or naturalistic (see Block 1986, 639; Von Eckardt 1993, 189-190).

---

[64] As suggested above (see note 53), and as will be shown in detail below (see Chp. Four, sec. 1.), there are two ways "representation bearers themselves" can determine their representational content. One is by virtue of their internal causal roles. Such a view, internal functional role semantics, is an option for advocates of classicism. The other is by virtue of their physical properties. In my view, the latter theory – structural isomorphism – is not naturally compatible with classicism (a claim I defend in Chp. Six).

It is my view that a TC and a TCD have been unwittingly conflated in recent theorising about mental representation. The thinking seems to be that because content is external or relational – since a description of content adverts to an object in the (represented) world – *then content determination must also be a matter of an externally-instituted relation.*[65] But this does not necessarily follow. Just because a representation bearer is about something beyond itself (and content is thereby external), does not mean content *determination* must be externalist. As suggested, this is to confuse a description of the representing relation (the content) with the determination of content (how the representation bearer came to have this content).

Again we have a confusion – namely, the TC/TCD conflation – brought about by an ambiguity in a key expression. The culprit, in this case, is "conditions of satisfaction". A description of a representation bearer's content can be thought of as a specification of its conditions of satisfaction; the condition in the (represented) world which would satisfy, or instantiate, what the representation bearer "says" of the world. Searle (1983, 13) points out that "conditions of satisfaction" is ambiguous between the *requirement* and the *thing required.* He writes:

> ...if I believe that it is raining then the conditions of satisfaction of my belief are that it should be the case *that it is raining* (requirement). That is what my belief requires in order that it be a true belief. And if my belief actually is a true belief then there will be a certain condition in the world, namely the condition *that it is raining* (thing required), which is the condition of satisfaction of my belief, that is, the condition in the world which actually satisfies my belief (1983, 13).

The way to interpret Searle's point is, I think, to understand the *requirement* sense of "conditions of satisfaction" as in some sense a feature of the intentional state itself. The *thing required* sense, on the other hand, is something external to the intentional state, a feature of the world. An earlier remark of Searle's suggests just such an interpretation: "...the conditions of satisfaction of the Intentional state are internal to the Intentional state....[P]art of what makes my wish that it were raining the wish it is, is that certain things will satisfy it and certain other things will not" (1983, 11). The *requirement* sense of "conditions of satisfaction" is what makes the

[65] Alternatively: since content is identified with the object to which a representation bearer is applied (on occasions of correct application), and application is external, then content determination must be externalist.

intentional state the type of intentional state it is (type-identifying, of course, by content). The requirement is constitutive of the representational content. The *thing required* sense of "conditions of satisfaction", on the other hand, is something beyond the content-identity of the intentional state. It is not constitutive of the representational content. Rather, it is the condition in the world itself which, if present, renders the intentional state (a belief, say) veridical.

Having a certain satisfaction condition is what it is for a state to be a *representational* state. A state's representational content specifies its conditions of satisfaction, in the "requirement" sense. That is, a representational state "says" certain things of the world; what it says is its conditions of satisfaction (requirement), whether or not such conditions actually obtain. Representational states, to put it in other terms, are semantically evalu*able* (Fodor 1987, 11). Beliefs can be true or false, perceptual states can be veridical or not. A certain actual state of affairs obtaining in the world is the "thing required" in order to satisfy the representational content. The "thing required" is what enables us to evalu*ate* the representational state.

A TC and a TCD are conflated, I think, because of a failure to notice the two senses of "conditions of satisfaction", and instead operate with only the "thing required" understanding. And this results in conceiving of intentionality as connectedness. To interpret "conditions of satisfaction" as the "thing required" is to see the required state of affairs in the world as constitutive of the representation bearer's content. It is then natural to think that some sort of externally-instituted relation between such an external state of affairs and the representation bearer *determines* its content. Because content is external, then content determination must be wide (externalist). But this is to run together a TC and a TCD. It is to see a specification of satisfaction conditions (in the "thing required" sense) – a specification of content – as a specification of what (by being in some relation to a representation bearer) determines the content. Moreover, if content determination is a matter of a relation between a representation bearer and its conditions of satisfaction in the "thing required" sense, the result is the sort of orthodox strategy described in Chapter Two (sec. 2.). That is, content determination viewed as an externally-determined relation between the "thing required" and a representation bearer gives *veridical* content. The TCD must then incorporate a statement specifying the privileged subclass of circumstance under which such a relation is content-determining, so that when the special circumstance does not obtain misrepresentation is the result.

The orthodox approach runs together a TC and a TCD such that the external object which is the "thing required" for veridical content is what – via some sort of externally-instituted relation to a representation bearer – determines content. Recognising the two senses of "conditions of satisfaction" means we can distinguish between a TC and a TCD, and drive a wedge between the issues of content determination and the veridicality of content. It can then be argued that just because a specification of the content mentions, or adverts to, some state of affairs in the external world which, were it to obtain, would mean that the content is veridical, does not mean such a state of affairs is constitutive of the content, or that it determines the content. A TC, then, describes representational content in the "requirement" sense of conditions of satisfaction. The "thing required" interpretation, on the other hand, pertains to the *veridicality* of content. In tokening a representation bearer, it is applied to (used to represent) some external object. But whether it is a case of veridical representation involves a comparison between the representation bearer's content and the external object to which the representation bearer is applied. If the "thing required" does obtain – or, more precisely, if the representation bearer is correctly applied – then, of course, we have veridical content. Conditions *external* to or beyond a representation bearer, that is, are relevant to an *assessment* of the veridicality of content. But such an assessment is an issue *beyond content determination.* A TCD can explain how content is determined without assuming that it must do so in an externalist fashion, and without being committed to the view that it is veridical content which is conferred upon representation bearers.

2. Two Approaches to Content Determination and Misrepresentation: A Summary

This section serves as a summary of the issues discussed in Chapter Two through to the present point. I will lay-out the commitments and assumptions of the orthodox approach to content determination and misrepresentation. Though these commitments are individually set-out, they should not be viewed as entirely independent. They are closely related, and it would be more accurate to describe them as ways of characterising the orthodox approach which emphasise slightly different aspects of the prevailing strategy.[66] Spelling-out the characteristics of the orthodox approach serves to highlight what is rejected by the alternative proposal. The scene will then be set for Chapter Four, where a TCD quite different to those discussed so far in this work will be developed.

The first assumption is that:

> (1) It is by virtue of being in some kind of relation to external objects which determines the content of representation-bearer types, and yields misrepresentation.

This, however, cannot be the case. As shown in the discussion of the crude versions of the causal and resemblance TsCD (Chp. Two, sec. 1.), attempts to get one relation to be both an explication of content determination and yield misrepresentation fail because there is just no room to manoeuvre misrepresentation into the picture. Yet while no one would advocate the crude causal theory, or the crude resemblance theory (the most graphic illustrations of the problem with assumption 1), the orthodox approach still does not break free of the assumption that one kind of relation between external objects and representation-bearer types both determines content and yields misrepresentation.

Fodor's "asymmetric dependence" proposal, of all the orthodox attempts to accommodate misrepresentation, perhaps comes closest to a rejection of assumption (1). Fodor recognises that the "false"-cause-to-symbol relation is (must be) somehow different to the "true"-cause-to-symbol relation. But rather than cutting the "false"-cause-to-symbol relation completely free, and, indeed, completely severing the determination of intentional object from the determination of content, Fodor goes only part of the way, suggesting that the "false"-cause-to-symbol relation hitches a ride on the "true"-cause-to-symbol relation.

Dretske's two proposals for accommodating misrepresentation (the learning period account, and the addition of a teleological component to the basic indicator story), and Millikan's purely teleological strategy, remain more clearly within the confines of assumption (1). Both theorists retain the idea that the one

[66] The reader will notice that the assumptions listed here do not straightforwardly line up with those put forward in Chapter Two, section 2. This is not an error. It is not my aim here to duplicate the earlier discussion. Now that we have a fuller picture of particular orthodox theories, I can provide a fuller picture of the orthodox approach in general.

type of relation both determines content and admits misrepresentation, but suggest that a principled distinction can be made between circumstances in which a type of representation bearer-representational object relation has content-determining powers, and circumstances in which such a relation lacks those powers. Fodor's proposal, even in the light of my comments above, can also be given such a characterisation. So, I arrive at the heart of the orthodox approach to content determination and misrepresentation (spelt-out in Chp. Two, sec. 2.), which is given by the second assumption.

> (2) The starting point for a TCD is the determination of veridical content. Propose a certain kind of naturalistic condition or special circumstance (condition C), under which content determination occurs and content is (supposedly) veridical. Then claim that conditions other than condition C are not content-determining, and tokened representation bearers in such circumstances are misrepresentations.

As I have shown (Chp. Two, sec. 2.), the theories of Fodor, Dretske, and Millikan differ in terms of what type of naturalistic condition, or special circumstance, is proposed as a (supposed) guarantee of veridical content.

The third assumption, or commitment, of the orthodox approach was also discussed in Chapter Two, section 2. It follows from assumption (2), and, in effect, is a slightly different way of viewing that assumption.

> (3) Accommodating misrepresentation, or providing an account of misrepresentation, is a required part of a TCD.

Because TsCD within the orthodox approach begin with veridical content determination, they must, in order to accommodate misrepresentation, say how conditions are different (non-content-conferring) when misrepresentation occurs. In order to let-in the possibility of misrepresentation, a TCD within the orthodox approach must provide an account of misrepresentation in the statement of the TCD itself. It must say, that is, how conditions are not content-determining when misrepresentation obtains. The account of misrepresentation, then, is a statement within the TCD itself in the sense that the TCD involves a distinction between Type I (condition C) and Type II (not C) situations.

The fourth characteristic of the orthodox approach relates to the focus on the tokening relation discussed in subsection 1.4. of the present chapter.

(4) In content-determining circumstances (condition C), the content of a representation-bearer type, and the object to which any token of the type is applied, are identical. That is, content and intentional object are determined by the same external object.

The orthodox approach begins with veridical content, and when content is veridical then, of course, the content of a representation bearer matches the object to which the representation bearer is applied. The orthodox approach, that is, proceeds by specifying a certain privileged subclass of representation-bearer-to-representational-object relation (condition C) which supposedly guarantees that content is veridical (correct application). Content determination is equated with correct application. Under condition C a representation-bearer type represents the external object with which it covaries (or with which it is supposed to covary). Now, as shown in subsection 1.4., in the perceptual case a representation bearer is applied to the object before the cognitive system which causes the bearer tokening. On the orthodox approach, then, what determines content also determines the representation bearer's intentional object. To accommodate misrepresentation, the approach must then say how content and intentional object *come apart*. And, again as shown above, the idea is that misrepresentational tokens happen when circumstances are other than those of condition C (assumption 2).

On the orthodox approach, content is determined by a representation-bearer type being in some sort of objective relation to the object it is correctly used to represent. This is what leads Cummins (1996, 29, 53) to call the TsCD within the orthodox approach "*use*" theories. Use theories have it that content is correct use. Content is identified with the external object confronting the cognitive system which causes, or covaries with, a representation bearer tokening. That is, since the external object confronting the cognitive system is the object the representation bearer is used to represent (the intentional object, or target), and veridical content is when content matches the object the representation bearer is used to represent, use theories see content determination as a matter of a relation between the intentional object and the representation bearer. Use theories are then forced into the strategy of making the condition C/not-C distinction. Condition C specifies a naturalistic condition under which use is supposedly guaranteed to be correct. But since not all uses are correct, it must be claimed that circumstances are sometimes different.

When condition C does not obtain, the external object a representation bearer is used to represent does not constitute its content, and misrepresentation results.

Cummins' (1996) labelling of TsCD within the orthodox approach as "use" theories is a succinct characterisation of the heart of the orthodox approach. The understanding of intentionality as connectedness considered in section 1., and the assumptions and commitments of the orthodox approach spelt-out in this section, can be ultimately characterised in terms of the idea that content is correct use. If content is correct use then the fundamental task in accommodating misrepresentation for the orthodox approach becomes one of specifying how content and intentional object *come apart*. At the heart of an alternative TCD, then, would be a rejection of the idea that content is determined in conditions of correct use. In section 1., I suggested that the intentionality-as-connectedness conception and the orthodox approach to content determination and misrepresentation results, at least in part, from a focus on the tokening relation. What is required of an alternative TCD is that content be determined by something other than a relation between the tokening object and a representation bearer on occasions of correct use. In section 1., I also suggested that it seems the only possibility is to claim that content is determined in an internalist fashion, by, that is, the representation bearers themselves. And at the heart of an alternative proposal for accommodating misrepresentation would be a rejection of the task of specifying how content and intentional object come apart. Instead, the reverse strategy is adopted: content and intentional object are *determined independently*. Sometimes the two will deviate, yielding misrepresentation; often they will match, thereby giving veridical representation.

To summarise. The alternative proposal rejects (1) above. The alternative proposal claims that the falsity of (1) means that content must be determined independently of that which determines the object to which a representation bearer is applied. In rejecting (2), the alternative proposal does not begin with veridical content. Rather, it builds the capacity for misrepresentation in from the start. Representation bearers have their content determined in a way *blind* to the veridicality of content. Veridicality is an issue *beyond* content determination. Veridicality is an issue concerning the match, or mismatch, between the representation bearer's independently determined content and intentional object. In building-in the capacity for misrepresentation, the alternative TCD does not require a statement *internal* to the TCD itself which elucidates misrepresentational conditions. This is its rejection of (3). And again because content and intentional object are determined independently, the alternative proposal rejects (4). Finally, though somewhat orthogonal to the central concern of misrepresentation, the alternative approach takes seriously the need to explain the selectivity of representation, discussed in subsection 1.3. above. It is time to develop a theory which can substantiate the alternative proposal.

CHAPTER FOUR

STRUCTURAL ISOMORPHISM AND THE DETERMINATION OF MENTAL CONTENT

Of all the resources available to construct a theory of mental content determination, resemblance is perhaps both the most maligned and undeveloped. So far in this work, a primary concern has been to show that the resemblance theory deserves more serious consideration than it is typically accorded. In Chapter One I sought to put the basis of the resemblance theory on firm footing by dispelling certain common misconceptions. Over the course of Chapters Two and Three, in a more competitive engagement with the orthodox TsCD, I argued for an alternative framework for the accommodation of misrepresentation. The alternative framework proposes that the capacity for misrepresentation derives from a representation bearer's content being determined independently of that to which it is applied. And this suggests the need for an internalist TCD, a theory where it is the mental representation bearers themselves which determine their representational contents. What is required, in other words, is some kind of resemblance TCD. Having motivated the task of subjecting the resemblance theory to a more detailed treatment, it's time to undertake the task itself. The focus of this chapter is the development and defence of the structural isomorphism version of the resemblance theory.

I begin the detailed development of the structural isomorphism TCD in the first section. The structural isomorphism TCD I formulate arises from consideration of various versions of the resemblance theory, including a variety which operates with a functional, rather than structural, conception of second-order isomorphism. I also distinguish the structural isomorphism TCD I recommend from other interpretations one finds in those rare discussions of structural isomorphism available in the literature.

In section 2., I sketch a certain view of mental content which naturally accompanies the structural isomorphism TCD, a fine-grained conception I call "microcontent". I touch upon some evidence for this view, as well as explaining its relationship to what was identified above (Chp. Three) as the selectivity of representation. Section 3. confronts what might be regarded as a powerful objection to the structural isomorphism TCD. The charge is that isomorphisms proliferate where they are not welcome, resulting in radical content indeterminacy. I show that there are various types of response available to the advocate of structural isomorphism. Finally, in section 4., I complete the examination into misrepresentation by tying up a few loose ends, including the issue of what determines targets or intentional objects.

1. The Varieties of Resemblance

As I've foreshadowed, there is not simply one resemblance theory but a cluster of related theories founded on certain varieties of resemblance. In this section I will distinguish between these differing versions of the resemblance theory. I set out to traverse somewhat tricky terrain that has not been well mapped, so let me first peg-out the logical geography.

In Chapter One I made a preliminary pass through the difference between first-order and second-order isomorphism, and my first undertaking will be to briefly revisit the distinction in order to rehearse the crucial elements of a second-order version of the resemblance theory (subsec. 1.1.). I then set about formulating the idea of a second-order resemblance theory in more detail (subsec. 1.2.). I take as my starting point – and, indeed, a reference point for much of the discussion of this section – an undeservedly neglected discussion by the cognitive psychologist Stephen Palmer (1978). I go on to reveal that there are *two* kinds of second-order isomorphism, *functional* and *structural* (subsec. 1.3.). In what might be a surprising result for many, I suggest that a functional rendering of second-order isomorphism corresponds to (internalist) functional role semantics. That is, functional role semantics, of the variety that appeals to the internal causal roles of representation bearers, is properly classed as a version of the resemblance theory. I argue that the structural kind of second-order isomorphism is to be preferred over the functional kind (subsec. 1.4.). The argument turns on the issue of the explanatory role of content. For the remainder of the section the focus is exclusively with various interpretations of *structural* isomorphism. I first distinguish Von Eckardt's reading of Palmer's account of structural isomorphism from my own interpretation of Palmer. The latter is then distinguished from the formulation I favour (subsec.

1.5.).[67]

1.1. Resemblance: First-order and Second-order Revisited

The resemblance theory is traditionally viewed as one which employs a first-order isomorphism between representation bearer and representational object. That is, a bearer represents its object by virtue of instantiating properties physically equivalent to those possessed by the object. However, as pointed out in Chapter One (subsec. 3.1.), a resemblance theory is not restricted to the notion of first-order isomorphism; a resemblance relation can hold in terms of a second-order isomorphism. As far as I can tell, the first expression of this distinction in the literature concerning mental representation is due to Shepard and Chipman (1970):

> ...the isomorphism should be sought – not in the first-order relation between a) an individual object, and b) its corresponding internal representation – but in the second-order relation between a) the relations among alternative external objects, and b) the relations among their corresponding internal representations. Thus, although the internal representation for a square need not itself be square, it should (whatever it is) at least have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of a persimmon (1970, 2).

The difference between first-order isomorphism and second-order isomorphism is that while first-order isomorphism involves a representation bearer having physical properties in common with its representational object, second-order isomorphism involves a systematic mirroring of the *relational* properties of a set of representational objects in the *relational* properties of a set of representation bearers. Variation among a set of representational objects comprises the relational structure of a represent*ed* "world". Variation among a set of representation bearers comprises the relational structure of a represent*ing* "world". It is *because* variation in a representing "world" systematically reflects variation in some "world" beyond

[67] Palmer's (1978) work is employed as a reference point for this section because it is the first detailed investigation into second-order isomorphism in general, and structural isomorphism in particular (though see Craik 1943 for an earlier *sketch* of structural isomorphism). In her survey of available TsCD, Von Eckardt (1993, Chp. 6.) views Palmer as the representative of structural isomorphism. Palmer's work, however, has received scant attention from philosophers of mind. Very recently, the philosopher Robert Cummins (1996) has come to endorse the structural isomorphism TCD, and it's fair to say that he is currently its most prominent advocate.

the representing world that the latter variations are represented by (are the contents of) the former variations. Represented and representing "worlds" are structures, or *systems* of representational objects and representation bearers. And content determination, on second-order isomorphism, is a matter of representing and represented systems having relational properties (relational structure) in common. [68]

The distinction between first-order and second-order isomorphism, then, turns on, respectively, whether representation bearers replicate physical properties of representational objects, or whether systems of representation bearers reproduce the relations holding among systems of representational objects. When the representing domain mirrors the relational structure of the represented domain, representation bearers need *not* have physical properties in common with the representational objects. The reproduction of relations means that a representing domain can be one which exhibits variation along a feature physically *non-equivalent* to a variable feature of the represented domain. In consequence, a second-order isomorphism can only be established in a "systemic" manner. That is, when physically non-equivalent properties are employed in a representational relation, a resemblance relation can only be secured by utilising a system of representation bearers and a system of representational objects. So representing and represented *systems* are *necessary* if the resemblance relation is not one of having physical features in common. If, on the other hand, the resemblance relation obtains by virtue of a sharing of physical properties, then it can hold between *either* an individual representation bearer and an individual representational object or a representing system and a represented system. A resemblance relation of physically equivalent properties, in other words, is *compatible* with both an "individual" and a "systemic" interpretation of that which does the representing and that which is represented.

Indeed, the "systemic" interpretation of first-order resemblance is, in my view, the right way to think about *non-mental* pictorial representation. Pictures are typically complex representation bearers, representing states of affairs containing a number of objects and features. As such it might be useful to think of them (and their like, diagrams, models, and maps) as comprised of, or instantiating, a number of representing systems. It can then be said that various first-order resemblance relations obtain between represented and representing *dimensions*. Variations in a

---

[68] I will shortly introduce some simple examples to illustrate the idea of second-order isomorphism.

dimension of the represented world, such as colour, are represented by corresponding variations in the physically equivalent dimension of the representing world. Similarly, spatial arrangement and proportionality of parts in the represented world are represented by the physically equivalent dimensions in the representing world. Such considerations suggest an attractive feature of the resemblance theory. Not only are certain objects and features represented, but also certain systematic relationships which hold between various objects, parts of objects, and features of the represented world can be captured. Indeed, often we get certain relations represented *for free*. If we have a map which represents the locations of three cities, then automatically we get a representation of, for example, their location in relation to one another, and the relative distances between them (see Haugeland 1987, 89-90). This is achieved because variations in a dimension of the represented world are represented, or tracked, by corresponding variations in a physically equivalent dimension of the representing world. Relations between items in the represented dimension are captured by differences in the "values" of items in the representing dimension.[69]

So, the significant difference between first- and second-order isomorphism

[69] There is a further attractive feature of formulating a first-order resemblance theory in terms of representing and represented systems. Such a systemic characterisation accords with a little-remarked but important characteristic of pictorial representation *in toto*. An advocate of the resemblance theory for non-mental pictorial content faces a complexity, perhaps even an objection, concerning differing *modes* of pictorial representation. 'Pictorial' representation is often taken to subsume, for example, crude monochromatic sketches, cartoon drawings, diagrams such as architectural floor-plans, black and white photographs, colour photographs, "realist" paintings, and complex three-dimensional models. As Neander (1987, 214-216) correctly recognises, the absence of *universal* respects upon which the resemblance between different modes of pictorial representation and representational objects holds is a difficulty confronting any attempt to construct an all-encompassing resemblance theory of pictorial content determination. The representing and represented systems conception suggests that a monolithic resemblance theory of pictorial content determination can only be formulated in very general terms. Any detailed formulation must specify, for each isolable mode of pictorial representation, the particular dimensions of the representing and represented systems between which the resemblance relation obtains. More "expressive" modes of pictorial representation (such as colour photographs) will, of course, have a higher number of such dimensions than their cruder relatives. And it's also the case that there are *intra*-dimensional differences between kinds of pictorial representation bearers in terms of their differing "resolution" (the range of different "values" they are capable of representing the dimension as exhibiting).

The systemic interpretation of resemblance also suggests that philosophers like Goodman (1969, 5), in objecting that any two paintings resemble one another more closely than any painting and its representational object, are operating with nothing more than a caricature of the resemblance theory. I said above (Chp. One, subsec. 3.1.), in responding to this objection, that pictorial representation bearers and their representational objects sustain a resemblance relation in terms of visual appearance properties; any version of the resemblance theory must specify the dimensions of the representing and represented systems between which the purported resemblance obtains. The considerations lately rehearsed show that anything approaching a satisfactory resemblance theory of pictorial content would need to specify the differing dimensions of resemblance involved in differing modes of pictorial representation.

is that between representation by a sharing of physical properties and representation by a sharing of relational structure. That established, I turn now to a more detailed examination of second-order isomorphism.

1.2. Second-order Isomorphism: Towards a More Detailed Account

I begin with Palmer's (1978) development of Shepard and Chipman's formulation. Palmer characterises a second-order isomorphism TCD as follows.

> The nature of representation [more precisely, content determination] is that there exists a correspondence (mapping) from objects in the represented world to objects in the representing world such that at least some relations in the represented world are structurally preserved in the representing world. In other words, if a represented relation, *R*, holds for ordered pairs of represented objects, (*x*, *y*), then the representational mapping requires that a corresponding relation, *R'*, holds for each corresponding pair of representing objects, (*x', y'*) (1978, 266-267).

For Palmer, the represented world and the representing world are comprised of a set of items and sets of relations that hold among these items. In the represented world, a given set of relations constitutes an aspect, or feature, of that world. In the representing world there exists a corresponding set of relations. A feature of the represented world, then, is represented by virtue of an isomorphism between the set of relations constituting that feature and a set of relations in the representing world.

Palmer (1978, 262-264) presents some simple examples in order to illustrate the idea of second-order isomorphism (see Figure 4.1.). In order to provide sufficient detail in the illustration of second-order isomorphism, it will be necessary to tacitly frame the discussion in terms of the two particular versions of second-order isomorphism, that is, functional and structural isomorphism. (I'll return to explicitly discuss the difference between the two in the next subsection). Consider a represented world consisting of four rectangles of differing heights and areas (world A in Figure 4.1.). Different aspects of such a world can be represented. For example, the relative height of the rectangles can be represented by the relative length of lines in the representing world. That is, the fact that rectangle *a* is taller than rectangle *b* is mirrored by the fact that line *a'* is longer than line *b'*. The relation *taller than* is represented by the relation *longer than* by virtue of an isomorphism which holds between the (*taller than*) relations among items in the represented world and the (*longer than*) relations among items in the representing world. Similarly, the relative *area* of the rectangles can be represented by the relative *lengths* of lines in the representing world (world D; Palmer calls it "larger" rather than "area"). Here a different aspect, or feature, of the represented world (area, rather than height) is being represented by line length. And the isomorphism is between the *greater area than* relations and *longer than* relations.

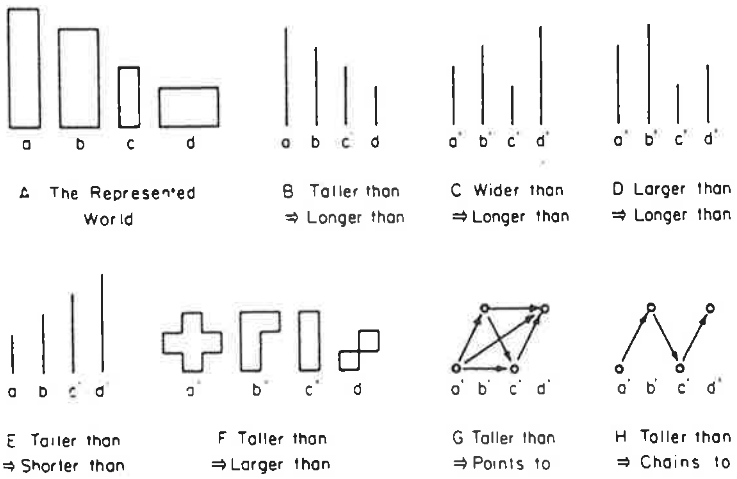

Figure 4.1. Some simple examples of second-order isomorphism. The represented world consists of the objects shown in A. The objects shown in B-H are different representing worlds. The expression beneath each of the representing worlds B-H indicates, first, the feature of A being represented, and, second, that feature of the representing world doing the representing. (From Palmer 1978, 263).

Of course, we are not restricted to a physical property such as line length to represent features of a represented world. On one way of securing a second-order isomorphism, any physical feature able to admit of systematic variations in correspondence with variations in a feature of the represented world can serve as a representing system in which items have their content by virtue of a second-order isomorphism to the represented world. So, we might employ, for example, brightness, density of cross-hatching, colour, number of dots, or voltage as the representing property. In such cases of representation by artefacts we are *constructing* a representing system so that variations therein mirror variations in a feature of the represented world, or we stipulate that variations in one system represent variations in another system. This involves an element of conventionalism. In order for items in the representing world to *operate as* representation bearers an interpreter must know what physical features are doing the representing, what physical features are being represented, and the "scale" of the mapping. Such a situation is clearly different, in two respects, from what is required for mental representation. First, the range of physical properties which can be posited as representing the represented world is, obviously, restricted to the sorts of properties actually possessed by the brain. In fact, *prima facie*, this might seem to constitute an objection to any attempt to construct a second-order resemblance TCD for mental representation (as noted in Chp. One, subsec. 3.1.). However, as I'll show over the course of the following chapters (see especially Chp. Six), this objection is without force. Second, a conventional element in the domain of mental representation would constitute a violation of the naturalistic constraint. Any requirement that a cognitive system itself literally know, or understand, the mapping function would saddle us with a regress; the explanation of intentional properties must bottom-out in nonintentional terms.

Consider now representing world G in Figure 4.1. Here the *taller than* relation is represented by the *points to* relation. The fact that *a*, for example, is taller than *b* in the represented world is preserved in the representing world G by *a'* pointing to *b'*. As Palmer indicates, representing world G constitutes a different way of representing *taller than* than representing worlds B-F. He writes:

> What makes this kind of representation different from those [representing worlds B-F] is that: (1) the representing objects corresponding to the rectangles are identical; and (2) new elements (the

> arrows) have been added that correspond explicitly to the relation being modeled (*taller than*). It is important to notice that the arrows of World G are not "object elements" but "relational elements". That is, the presence of a given arrow in the [representing] world does not correspond to any single object in the represented world. ... The difference [between representing world B, for example, and representing world G] is that although *longer than* is a relation that can hold between an ordered pair of objects, *is arrow-connected to* is a relation that can hold between an ordered pair of objects only by virtue of each being related to a third element (the relational arrow) in a particular way (1978, 263-264).

Palmer's account of the difference between representing worlds B-F and representing world G can be summed-up as follows. Representing world G lacks inherent structure. The isomorphism to represented world A is not secured by the inherent structure of the representation bearers in representing world G. Rather, the structure obtaining in representing world G (and thereby the isomorphism) must be externally *imposed* on the system of representation bearers. In contrast to representing worlds B-F (and the characterisation of second-order isomorphism provided in the preceding two paragraphs), then, the account of representing world G (and thereby the characterisation of second-order isomorphism provided in this paragraph) concerns a *functional*, rather than *structural*, kind of second-order isomorphism. A functional isomorphism implicates the *roles* of representation bearers in the representing system. With structural isomorphism, in contrast, the relations between representation bearers are *physical* rather than functional. I'll return to the difference between these two versions of second-order isomorphism in the next subsection.

A second-order isomorphism TCD can be characterised in more formal – and *generic* – terms as follows. Represented and representing worlds are comprised of a set of items and sets of relations holding among these items. A mapping, or correspondence, between the two worlds holds such that: (1) for each item in the represented world there is a corresponding item in the representing world; (2) for each relation in the represented world there is a corresponding relation in the representing world; and (3) whenever a relation in the represented world holds of an n-tuple of items, a corresponding relation in the representing world holds on the corresponding n-tuple of items (see Cummins 1996, 96; McGinn 1989, 179-180;

Palmer 1978, 294; Von Eckardt 1993, 207-209).[70] And as Cummins (1996, 96) correctly points out, it is important to realise that the representing world does not represent the represented world because the items and relations in the former represent items and relations in the latter. Rather, items and relations in the representing world represent items and relations in the represented world *because an isomorphism* (or set of isomorphisms) obtains between the two worlds. The items and relations which constitute the representing world are contentful (in part[71]) because they are elements of a representing structure, or system. This might seem to give rise to global content holism, an implication abetted by talking of represented and representing "worlds". However, *global* holism – the view that the content of any item in the *entire* representing world (a complete cognitive agent, for example) depends on its relations to every other contentful item in the entire representing world – does not follow from the commitment to represented and representing structures. Rather, the representing structure ranges over items and relations in a *narrowly* circumscribed represented domain (of which much more on below; see sections 2. and 3. of this chapter).

1.3. Two Kinds of Second-order Isomorphism: Functional and Structural

So far in this work, I've portrayed the resemblance theory as an alternative to the existing or prevailing TsCD, a "new" account of content determination (in the sense specified in the Introduction, note 3.). What I've called the "orthodox" theories, for example, assert that content determination is a matter of an externally-instituted relation between environmental object and internal representation bearer, not something concerning the "representation bearers themselves" (in particular, their relations to one another). Orthodox TsCD do not – in contrast to the resemblance theory – posit an internally-structured system of representation bearers which, by virtue of an isomorphism to a system of representational objects, does the content-

---

[70] This is a characterisation of *exhaustive* representation; that is, every item, feature, and relation in the represented world is captured in the representing world. The representing world in such a case is what McGinn (1989, 180, note. 15) calls a "*total* model". Mental representation is not like this (neither is much non-mental representation, for that matter). As I suggested above (Chp. Three, subsec. 1.3.), *selectivity* is a characteristic of mental representation, and the TCD developed in this work aims to respect this characteristic. I'll have more to say on selectivity in the next section.

[71] I say "in part" because, as argued in Chapter One (sec. 2.), there is more to an item's representational status than simply bearing some kind of relation to the object represented.

determining work. I now need to qualify my initial way of marking-out the content-determination landscape.

The qualification has been foreshadowed. I've already pointed out that there is not one resemblance theory, but various versions founded on differing kinds of resemblance; distinguishing between these various versions is, of course, the main task of the current section. More particularly, I've suggested that there are two varieties of second-order isomorphism, two ways "representation bearers themselves" determine their contents. I now propose to flesh-out this claim. The upshot is that only one variety of second-order isomorphism issues in a *distinct* TCD, a genuine alternative to *already existing* or prevailing accounts of content determination.

First, a representing system can be second-order isomorphic to a represented system in terms of *physical* relations among the representation bearers. The representing system, that is, consists of a physically structured network of representation bearers. Second, a representing system can be second-order isomorphic to a represented system in terms of *causal* relations among the representation bearers. The representing system, that is, consists of a causal network of representation bearers. The latter, I claim, corresponds to internalist functional role semantics (FRS), a TCD which has been around for some time, and one which has attracted its fair share of adherents. On such a view, causal relations among a network of representation bearers mirror some network of relations among a system of representational objects. This kind of second-order isomorphism TCD, then, operates with a *functional* isomorphism. The former, of course, is the TCD being proposed in this work, and it is this theory which does constitute a *novel* TCD. On such a view, physical relations among a network of representation bearers mirror some network of relations among a system of representational objects. This kind of second-order isomorphism TCD, then, operates with a *structural* isomorphism. I've called it "structural" isomorphism because it involves an isomorphism between the relational physical structure of a system of representation bearers and the relational physical structure of a system of representational objects. "Physical" isomorphism would not really do because first-order isomorphism it to be characterised in terms of a sharing of physical properties.

Certain characterisations of second-order isomorphism provided above (for example, in the previous subsection, those contained in the Shepard and Chipman quotation and the "more formal" account) are compatible with both functional and structural varieties of second-order isomorphism. The idea that second-order isomorphism comes in a functional form is something that has gone largely unrecognised; (internalist[72]) FRS is not usually viewed as a version of the resemblance theory. I'll return to specify the sense in which it is in a moment. First, however, I want to further pursue the two kinds of second-order isomorphism by returning to my reference point, namely, Palmer's (1978) discussion.

Though Palmer does not line up the idea of functional isomorphism with FRS (FRS was only just being conceived at the time of Palmer's writing), he clearly recognises that second-order isomorphism admits of both a functional and structural interpretation. He begins by showing that the initial formulations of second-order isomorphism can be seen to fit with a *propositional* or linguiform style of representation. Recall his initial formulation of second-order isomorphism:

> ...there exists a correspondence (mapping) from objects in the represented world to objects in the representing world such that at least some relations in the represented world are structurally preserved in the representing world (1978, 266-267).[73]

Palmer goes on to point out that such a characterisation requires that a representing world contain items which correspond to represented items. All that is required for representing a certain relation which holds between items in the represented world is that items in the representing world are in some relationship which *"functionally corresponds"* to the relation in the represented world (1978, 294-295, emphasis added). Propositional representational systems achieve such a functional correspondence by having *separate* relational items. These relational items represent relations by themselves being related to object items. Linguiform or sentential representation is, of course, a clear case of this way of attaining functional correspondence. Palmer writes:

> Words referring to objects are related in syntactically ordered strings through relational words (verbs, prepositions, and the like). The sentence, "The ball is under the table", specifies a relationship between

[72] From now on I'll omit the "internalist" qualifier, on the understanding that it is the variety of FRS which appeals to the internal causal roles of representation bearers to which I'm referring.

[73] The use of "structurally" here might seem to suggest that Palmer is referring to structural isomorphism in particular, not second-order isomorphism in general. However, it is clear from the context of Palmer's discussion that he intends "structural" preservation of relations in the representing world to be *neutral* between what I've called "structural" and "functional" means of securing a second-order isomorphism. See also note 29. of Chapter One above.

> the ball and the table that can only be understood by virtue of their syntactic relationships to the relational construction "is under" (1978, 295).

So, a second-order isomorphism can be established in terms of the *role* of symbolic representation bearers. Relations among representational objects, that is, can be mirrored by the representing system in a functional manner, a way compatible with a language-like style of representing. As so formulated, second-order isomorphism doesn't provide for a TCD significantly different from already existing TsCD. We require something other than a functional means of mirroring relations among a system of representational objects.

Palmer's (1978, 270-271, 296-297) proposal sees him stepping into the controversy over the propositional/analogue distinction, and differentiating *"extrinsic"* from *"intrinsic"* methods of preserving relational structure. With extrinsic representation, the structure of a represented relation is preserved by relations between representation bearers which inherently lack this structure; the structure of the relation is imposed by the relation being represented. Recall the above discussion of how representing world G differs from representing worlds B-F in Palmer's examples (subsec. 1.2.). The system of representation bearers comprising representing world G does not have inherent structure. When one representation bearer *points to* another (and thereby represents the fact that one rectangle in represented world A is *taller than* another), there is nothing to prevent the latter simultaneously *pointing to* the former. The structure exhibited by representing world G so as to mirror the structure of its represented world is externally imposed. Palmer sees propositions as extrinsic representations, and writes:

> ...the significance of using relational elements in representing relations is that propositions are extrinsic representations. The reason is that any object can, in principle, be connected to any relational element in any fashion. Hooking up object elements by relations to relational elements places no constraints whatsoever on the nature of the relations represented. Thus, *whatever structure there is in a propositional representation exists solely by virtue of the extrinsic constraints placed on it by the truth-preserving informational correspondence with the represented world* (1978, 296).

With extrinsic representation, structure must be forced into the representing relation in such a way that it accords with the structure in the represented relation. The point here is that the representation bearers which are employed to represent the relations between representational objects don't themselves inherently possess physical properties which systematically mirror the representational objects in the appropriate way. That is, an extrinsic means of preserving relational structure – a functional isomorphism – cannot be established by the symbolic representation bearers *themselves*. The intrinsic physical properties of symbols bear an arbitrary relationship to their representational objects. The structure exhibited by a system of symbols is not that of inherent physical structure. Consequently, the symbolic representation bearers must be disciplined by some other means. To secure a functional isomorphism, the roles of the symbols must be determined by something other than the symbols themselves, a set of "rules" of some kind.

With intrinsic representation, in contrast, a representing relation has the same inherent structure as its represented relation. The inherent structure is found in the physical properties of the representation bearers comprising the representing world. Recall representing world B in Palmer's examples (subsec. 1.2.). The important point here is that the representing relation *longer than* possesses the *same inherent constraints* as the represented relation *taller than*. That is, when one representation bearer is longer than another, then the second cannot be simultaneously longer than the first (just as in the case of the *taller than* relation). In the case of intrinsic representation, separate relational elements are not employed. Rather, because of the natural correspondence between the inherent structures of the represented and representing worlds, relationships among represented objects are mirrored by relationships among representing items. Palmer contends that analogue representation can be defined in terms of intrinsic representation: "*...whatever structure is present in an analog representation exists by virtue of the inherent constraints within the representing world itself, without reference to the represented world*" (1978, 297). Because extrinsic representations have no inherent structure, they can be made to represent many different kinds of relations. This is not the case for intrinsic representations, where the inherent structure of the representing relations tightly constrains the kinds of relations which can be represented. Palmer (1978, 297)) terms the method of mirroring relational structure in the inherent structure of the representing world *"natural"* isomorphism. Natural isomorphism is more abstract than first-order isomorphism because the representation bearers comprising the representing world need not have properties which are physically equivalent to those possessed by the representational objects comprising the represented world. It is, however, more concrete than functional isomorphism, because the latter specifies only that relational structure be functionally captured, and this, as just shown, can be done in an extrinsic fashion.[74]

It turns out, then, that there are two general kinds of second-order isomorphism. Palmer's "natural" isomorphism (which implicates "intrinsic" representation) is what I call "structural isomorphism", and it constitutes a novel TCD (in the sense that it differs from already existing TsCD). Structural isomorphism is founded upon the relational *physical* structure of a representing system. Functional isomorphism (which implicates "extrinsic" representation) grounds FRS. Here it is the relational *causal* structure of a representing system that matters. It is not always noted that FRS utilises an isomorphism, though sophisticated discussions (Fodor 1987, Field 1978) make clear the appeal to such an idea. The starting-point of FRS is, of course, functional role. The causal relations between mental representation bearers constitute a causal network. The functional role of a representation bearer is its location in such a causal network. According to FRS, the determination of content is to be explicated in terms of the causal interrelations among representation bearers. This, however, does not yet answer the question of content determination, "the hard question" as Fodor terms it: "How does locating a mental state in a causal network determine its intentional content?" (1987, 78). What is it about the causal role of a mental representation bearer that confers its content? It is here that isomorphism enters the picture. The idea is that there is an isomorphism between the network produced by the *causal* relations among mental representation bearers and the network produced by the *semantic* relations among the objects of such representation bearers (the objects being, on standard versions of FRS, propositions). The isomorphism, then, is such that the causal role of a representation bearer mirrors the semantic role of the proposition

---

[74] On the basis of the above discussion, I think that extrinsic representation, functional isomorphism, is compatible with conventional computation, and thereby the *classical* computational theory of mind. In conventional computational systems, a set of computational rules determines the roles of the symbolic representational items. In contrast, "intrinsic" representation or structural isomorphism is not compatible with classicism. A structural isomorphism, again as shown above, obtains when the inherent relational physical structure of a system of representation bearers corresponds to the relational physical structure of a represented domain. While incompatible with classicism, the intrinsic means of preserving relational structure is, in my view, compatible with the *connectionist* computational theory of mind. I discuss this kind of point in more detail in Chapter Six below.

that is its object (Fodor 1987, 77-79). Since FRS employs a second-order isomorphism between representing and represented domains, it is properly classified as a version of the resemblance theory.[75]

It should be clear, however, that FRS differs from the structural isomorphism TCD I am pursuing in this work. As I've made plain, for FRS the isomorphism is between sets of causal relations among representation bearers and sets of semantic relations among the objects of representation bearers. With the structural isomorphism TCD, in contrast, the isomorphism is between physical relations among representation bearers and certain physical relations among items in the represented world. I'll return to flesh-out structural isomorphism in a little more detail shortly. I next want to consider an issue that suggests the structural variety of second-order isomorphism is to be preferred over the functional kind.

1.4. Functional Versus Structural Isomorphism: The Explanatory Role of Content

In this subsection I propose to briefly consider the explanatory role of content, an issue I first introduced in the critical examination of Dretske's (more recent) account of content determination (Chp. Two, subsec. 2.3.). Specifically, the *desideratum* that representational content be of genuine explanatory relevance, I will suggest, strongly mitigates against a TCD founded on functional isomorphism, and in favour of the structural isomorphism TCD. That is, I will argue that a functional isomorphism renders content-explanation vacuous. In contrast, on structural isomorphism representational content has straightforward and robust explanatory significance. It might be thought that such a task is largely superfluous. After all, I've explicitly set out in this work to develop the structural isomorphism TCD. And as I remarked in the Introduction, structural isomorphism is worthy of

[75] FRS is a broader church than the standard interpretation presented here. It can be formulated in terms of a second-order resemblance relation between a set of causally related representation bearers and a set of semantically, logically, or causally related representational objects. It's worthwhile pointing out that Von Eckardt (1993, 206, 208) also classifies FRS as a kind of resemblance TCD. Like Fodor (1984/1990a, 33), Von Eckardt claims that all available TsCD are ultimately founded on either resemblance or causality (or both; Von Eckardt claims that Millikan's teleosemantics looks to have elements of both similarity and causality, though I won't pursue this further here). It's important to recognise that TsCD are to be classified as resemblance, on the one hand, or causal, on the other, in terms of the nature of the relation between representation bearers and representational objects. FRS, of course, makes use of the causal relations which hold among representation bearers, but this is not the content-determining relation. The nature of the relations that hold *within* the *representing* system is the criterion by which we demarcate the functional version of a second-order resemblance theory from the structural version.

investigation in its own right. Most generally, given that the naturalisation of intentionality project has reached something of a stalemate, it's important that no area of the content-determination landscape remain unexplored, and the structural isomorphism TCD has received little in the way of detailed development. However, having revealed that there are two varieties of second-order resemblance theory, I think it's important to at least touch upon a reason for thinking that the structural version is to be preferred. I should point out that I'm not going to wade into the murky and deep waters of the causal role of content debate; the literature is large and a proper examination would take me beyond the central concerns of this work. Rather, I simply intend to clarify why functionalism about content determination undermines the explanatory appeal to representational content.

It is often noted that a functionalist approach to content determination trivialises, or renders problematic, the explanatory role of content (see, for example, Block 1986, Fodor 1990c, and especially Cummins 1996). It is not difficult to see why. The central idea of FRS, as discussed in the previous subsection, is that there is a causal network generated by the interrelations among representational objects and a causal network generated by interrelations among representation bearers. Representation bearers have their content conferred by virtue of a second-order isomorphism between the two networks. The two networks are produced by the causal roles of their constituents. In particular, the representation bearers making-up the representing network only have their contents *ultimately* determined by means of their cognitive *consequences* in the cognitive system harbouring the network. And, *prima facie*, this clearly looks to be in tension with the view that representational content is in some way relevant to – in some fashion a determinant of – the particular cognitive and behavioural effects had by representation bearers. Here's how Cummins puts the point, framed in terms of conceptual role semantics (CRS), a species of FRS:

> According to CRS, talk of an attitude's content is just shorthand for, among other things, talk of the consequences of its occurrence. So, of course, CRS won't allow you to explain the consequences of the occurrence of an attitude by appeal to that attitude's content.
>
> ...
>
> Content, on this view, is really a kind of fiction, for it treats talk of

> content as just a convenient way of referring to the epistemic liaisons enabled by a representation without having to mention them explicitly....Once we realize this, it is obvious that CRS isn't going to allow us to explain behavior by appeal to content....CRS trivializes content explanations (1996, 39).

Those TsCD which, like CRS and FRS, are premised on the roles of representation bearers are manifestly, even self consciously, "use" theories. And if content is in some way determined by use, then, to put it charitably, it's difficult to see how content could be explanatorily potent, since that means that content in some fashion influences use.

We can gain a useful perspective on how functionalist accounts of content determination leach the explanatory power from representational content by focusing on the objective relation between representation bearers and representational objects. Indeed, I now intend to suggest that, like the fundamental point of difference between original and derived intentionality argued for in Chapter One (subsec. 2.3.), the issue of the explanatory role of content *ultimately* comes down to something about the objective representation bearer-representational object relation. The upshot is an illuminating division *within* the doctrine of original intentionality. I claimed that the difference between original and derived intentionality ultimately turns on, respectively, the presence or absence of some kind of objective relation between mental representation bearers and representational objects. To posit an objective relation between mental states and the objects they represent is to hold that mental states are contentful in their own right; mental content is not derived from the interpretative interests of other cognitive agents. Furthermore, there is a fact of the matter beyond the interpretive interests of cognitive agents as to the real, determinate content of mental states. Derived intentionality, in denying the existence of an objective bearer-object relation, holds that intrinsic and determinate content is a myth. The difference between a TCD of the sort grounded in functional isomorphism (which looks to make content-explanations trivial) and a TCD like structural isomorphism, I believe, likewise comes down to something about the nature of the relation between representing and represented systems. In this case, the issue is how the objective relation is put in place, or secured. On a *functional* isomorphism TCD, *the objective, content-determining relation is only ultimately fixed or put in place by* *means of the effects of representation bearers.* A functional isomorphism, in other words, is *not* in place antecedent to, or independently of, the causal consequences of representation bearers. The objective relation is only secured by means of the causal roles of representation bearers, the effects of representation bearers, of course, being a constituent of such causal roles. Content determination is a consequence of cognitive effects. And if the content-determining relation is only ultimately established by the effects of representation bearers, then patently it's very difficult to see how content could straightforwardly be accorded genuine explanatory relevance. On the *structural* isomorphism TCD, in contrast, the objective, content-determining relation is secured by the physical properties of representation bearers, not their effects or causal roles. That is, a structural isomorphism is put in place antecedent to the cognitive consequences of participating representation bearers. The structural isomorphism TCD completely decouples the establishment of an isomorphism, and thereby the determination of content, from the causal relations into which representation bearers enter (in terms of both the tokening relation, and the internal causal relations which constitute cognitive processes and, eventually, drive behavioural outputs).

In the first place, then, there is the fundamental divide between original and derived intentionality, in terms of, respectively, the postulation or denial of an objective relation between mental states and their representational objects. On derived intentionality, because there is no objective relation which fixes content antecedent to the cognitive and behavioural consequences of representation bearer tokenings, the explanatory role of content is at best obscure, if not vacuous. But original intentionality does not necessarily accord straightforward explanatory significance to content. *Within* the doctrine of original intentionality there is another fundamental division, that between a kind of TCD where the objective bearer-object relation obtains by virtue of the effects of representation bearers, and the kind where the objective relation holds independently of and antecedent to the cognitive consequences of representation bearers. This way of carving the space of possible TsCD has the two varieties of second-order resemblance theory in different camps. A functional isomorphism is ultimately a *consequence* of cognitive processing. A structural isomorphism is secured independently of cognitive processing, and content can thereby influence the *course* of cognitive processing. The explanatory relevance of content straightforwardly follows from structural, but not functional, isomorphism, and for this reason the structural isomorphism TCD possesses a clear advantage.

As Cummins (1996, 40) rightfully notes, many advocates of a functionalist TCD won't be terribly worried by this, since many functionalists hold that content is explanatorily vacuous anyway. But I wonder if functionalists cleave to the impoverished nature of content-explanations on independent grounds, or because it looks to be rather obviously entailed by the TCD they defend? For the most part, I suspect the latter is in play. And if the explanatory inertness of content is a consequence of cleaving to a functionalist TCD, then a TCD which doesn't have such a consequence wins hands down. Indeed, Cummins (1996, 41) goes so far as to suggest that the successful formulation of a TCD which accords a robust role to content-explanations is enough to refute the explanatory inefficacy of content, and thereby those TsCD which look to entail it. I can pretty much agree with such a sentiment. Functionalism about content determination is self-defeating; it erodes the whole purpose of engaging in a representationalist understanding of the mind. Representationalism should explain cognition; cognitive processing doesn't explain representation, in particular, content determination. Buying representationalism buys you enormous explanatory power. The *very point* of proposing representationalism in the first place is that representational content is justifiably thought to be relevant to the explanation of cognitive and behavioural performance, because physically-instantiated representational content in large part determines cognitive and behavioural performance. And not just successful and appropriate performance. As Cummins (1996, 41-50) correctly highlights, we need content to have explanatory significance so as to account for inappropriate behaviour or performance failure: misrepresentation explains misbehaving. Misrepresentation, as I've repeatedly emphasised, involves a distinction between the use or application of a representation bearer, and the bearer's content. Functionalist-type TsCD are particularly stark examples of the difficulties use theories face in accommodating misrepresentation. We need the content/use distinction to make sense of the idea that a cognitive system can misuse a representation bearer. Use theories manifestly deny this distinction. What looks to be deviant use of a representation bearer with a given content is, on FRS, reinterpreted as correct use of a representation bearer with a different content assignment, since content is determined by the actual use of representation bearers. As Cummins (1996, 41) puts it, "any evidence of representational error is better evidence for a different assignment of contents." Furthermore, we need the content/use distinction to provide for the distinction between representational error and reasoning (or processing) error (see Cummins 1996, 50). A cognitive system can get things wrong by misrepresenting the true situation, or by erroneous or inappropriate processing of representation bearers. On a functionalist approach to content determination, these two independently variable factors are conflated.

Functional and structural isomorphism sit on different sides of the fence when it comes to the explanatory role of content. While the explanatory significance of content looks, at best, problematic on a functionalist TCD, it finds a natural home with structural isomorphism. It would be interesting to enter the structural isomorphism TCD into the established and ongoing debate over the nature and viability of content-explanations in a science of the mind. But as regards the concerns of this work, the above preliminary discussion must suffice; it would be inappropriate to devote considerable space to the implications of the structural isomorphism TCD in the absence of its adequate formulation and defence. So it is to the task of elaborating upon the structural isomorphism theory that I now return.

1.5. Systemic Structural Isomorphism

Over the past two subsections I've made the important distinction between a structural and a functional variety of second-order isomorphism, and argued in favour of the structural isomorphism version of the resemblance theory. However, there are not just different versions of the resemblance theory (first-order, second-order structural, and second-order functional), there are differing possible interpretations and formulations of the structural isomorphism TCD. By way of further specifying the structural isomorphism TCD I wish to recommend, I now propose to examine these possible alternative formulations. I will begin with Von Eckardt's (1993) interpretation of Palmer's (1978) presentation. Von Eckardt's discussion is the only substantive philosophical examination of Palmer's proposal of which I am aware, so it earns a place in a work dealing with the detailed development and defence of the structural isomorphism TCD. A central concern of Von Eckardt's discussion is to distinguish Palmer's structural isomorphism TCD from FRS. But rather than focussing on the difference argued for above (subsec.

1.3.), she proposes another point of departure.[76] I will argue, however, that the version of structural isomorphism Von Eckardt ascribes to Palmer is not the only way of rendering a structural isomorphism TCD. I will then compare my own interpretation of Palmer with the formulation of structural isomorphism I favour. The difference here is subtle, though not insignificant. Finally, and by way of further clarifying my formulation, I return to the issue of whether Von Eckardt has interpreted Palmer correctly. The suggestion will be that Von Eckardt's characterisation of structural isomorphism is not particularly plausible.

Von Eckardt sees Palmer as the representative of the structural isomorphism TCD, and frames her discussion accordingly. Consider, then, her characterisation of Palmer's theory:

> A representation bearer RB represents a representational object O under some aspect D if there exists a set *G* of relations which constitute RB and a set *D* of relations which constitute O such that *G* is isomorphic to *D* (1993, 207).

Von Eckardt sees Palmer's theory as one where an isomorphism obtains between a set of relations *constituting* the representation bearer (the representing world) and a set of relations *constituting* the representational object (the represented world). And, for Von Eckardt, it is this which serves to distinguish Palmer's structural isomorphism theory from FRS, as the following passage makes clear.

> The structural isomorphism approach makes use of an isomorphism between the structure of a representation bearer and the structure of the representational object. The functional role approach also relies on the existence of an isomorphism. However, in contrast with the structural isomorphism approach, the isomorphism in question is between sets of relations that hold *among* the relevant representation bearers and representational objects, respectively, rather than between sets of relations that *constitute* the individual representation bearers and objects (1993, 210-211).

According to Von Eckardt, then, on Palmer's theory a representation bearer is made up of a certain set of relations in the representing world. Similarly, a representational object is made up of a certain set of relations in the represented

[76] Indeed, Von Eckardt (1993, 206-214) glosses over the difference between the structural isomorphism TCD and FRS I described above (subsec. 1.3.). That is, it is never explicitly stated that the relations which hold among representation bearers are *physical* in the case of structural isomorphism, and *causal* in the case of FRS. To my mind, this is a rather significant oversight.

world. Recall the above example from Palmer where the *taller than* relation among rectangles is represented by the *longer than* relation among lines (subsec. 1.2.). Von Eckardt sees the representational object as the set of *taller than* relations that holds for each pair of rectangles, and the representation bearer as the set of *longer than* relations that holds for each pair of lines. Now, my primary concern at this moment is not with the issue of whether Von Eckardt has interpreted Palmer correctly, but, rather, lies with the question as to whether Von Eckardt's interpretation of Palmer's theory is the only possible rendering of a structural isomorphism TCD. I'll return to the question of whether Von Eckardt has got Palmer right subsequently.

The short answer is that a structural isomorphism TCD need not follow Von Eckardt's interpretation of Palmer's position. According to Von Eckardt, as just explained, Palmer's theory is one where a certain set of relations *constitute* the representation bearer and another set of relations *constitute* the representational object, and an isomorphism obtains between the two sets of relations. On Von Eckardt's interpretation, individual representation bearers and representational objects are *structures*. The structure of a representation bearer, or representational object, consists of the set of relations which comprise the representation bearer, or representational object. And the isomorphism obtains between the structure of the representation bearer and the structure of the representational object.

However, it is possible to propose a structural isomorphism TCD where, like FRS, the isomorphism is between sets of relations which hold *among* certain representation bearers and sets of relations which hold *among* certain representational objects. On this alternative proposal, representation bearers and representational objects are individual items that belong to a *system* of representation bearers and representational objects. The relevant structure, in this case, is that of the relations which hold among representation bearers, and representational objects. And the isomorphism obtains between the structure within the system of representation bearers and the structure within the system of representational objects. Given that the isomorphism is between representing and represented system, the theory could be dubbed "*systemic* structural isomorphism". For the sake of brevity, however, I'll stick to "structural isomorphism". Whenever I subsequently speak of the structural isomorphism theory "I favour" or "being developed in this work" (or words to that effect), it is the systemic version just described to which I am referring. Of course, it will not have gone unnoticed that, even prior to the introduction of the distinction between two kinds of second-order isomorphism, and the systemic account of representing system lately emphasised, I have typically portrayed the structural isomorphism TCD along the lines rehearsed here. The systemic account of the structural isomorphism TCD solidifying at this stage, then, should not come as a surprise.

By way of elaborating upon the idea of the structural isomorphism theory proposed above, let me now consider it in regard to Palmer's position. Here I'm going to distinguish between Palmer's theory as *I* (not Von Eckardt) interpret it and the version of structural isomorphism I ultimately favour. Palmer writes primarily of *relations* being represented. Recall his simple examples where various sorts of relations among items in the represented world are represented by physical relations among items in the representing worlds. For example, the *taller than* relation is represented by the *longer than* relation (in World B), the *wider than* relation is represented by the *longer than* relation (in World C), and the *greater area than* relation (Palmer calls it "larger than") is represented by the *longer than* relation (in World D). Now, Palmer writes:

> The representing worlds B, C, and D show how different aspects of the same represented world could be modelled by the same representing world [that is, a world which does the representing by employing the same kind of physical property, namely, line length]. World B reflects the *relative height* of the rectangles (*a, b, c, d*) by the relative length of the corresponding lines (*a', b', c', d'*). In other words, the fact that *a* is *taller than b* in World A is reflected by the fact that *a'* is *longer than b'* in World B. Similar statements can be made for *any pair* of rectangles in World A and corresponding lines in World B.
>
> ...
>
> The implication is that any question that could be answered about *relative height* in A could be equally well answered by considering relative length in B as long as the mapping of rectangles to lines were known.
>
> World C reflects the *relative width* of rectangles in A by the relative length of lines. For example, the fact that *d* is wider *than any other rectangle* in World A is reflected in the fact that *d'* is *longer than any other* line in World C (1978, 262-263; emphasis added).

My reason for quoting Palmer at some length here is that his emphasis on the representation of *relations* can somewhat obscure something fundamental to the idea of structural isomorphism. What is left implicit in Palmer's discussion is that it is not just that *d'* is longer than *any other* line in world C which represents the fact that *d'* is wider than *any other* rectangle in World A. In addition, *d'* is longer than *c'*, for example, by an *amount* that corresponds to the *amount* by which *d* is wider than *c* (relative, of course, to the magnitude scales being employed in the representing and represented worlds). Similarly, *d'* is longer than *b'* by an amount that corresponds to the amount by which *d* is wider than *b*. If the mapping function is known, then not only can we answer questions concerning, for example, whether *d* is wider than *c*, and questions concerning to what extent *d* is wider than *c*, we can also answer questions about the *absolute width* of items in the represented world. That is, any line in the representing world (World C) has a *value* constituted by its length which corresponds to a *value* of a variable (width) in the represented world (World A). And, turning from the above non-mental examples to mental representation, if the mapping function is "instantiated", then a cognitive system can *represent* the values of a certain variable or feature of the represented world. Each of Palmer's "representing worlds" is what I've been calling a "representing system".

A represent*ing* system represents a rather specific feature, dimension, or variable of the represented world. Sophisticated representational devices (including, at the upper end, cognitive creatures such as ourselves) contain very many representing systems, each of which represents a specific feature of the represented world. What I've been calling a "represent*ed* system" is a feature or variable of the represented world, and consists of the values the feature or variable can exhibit. Palmer's "represented world" is, then, comprised of many different represented systems. (At the upper end it is the set of everything that can possibly be represented, including mental states themselves, since a mental state can be about another mental state). The isomorphism obtains between the possible values of the variable comprising the represented system and the possible values of the variable comprising the representing system. The difference in magnitude between any possible values in the represented system is mirrored by the difference in magnitude between any possible values in the representing system. I arrive, then, at a characterisation of structural isomorphism that was foreshadowed early in this work: variations in the value of a feature, dimension, or variable which comprise the represented system are tracked by variations in the value of a feature, dimension, or variable which comprise the representing system.

It is the above characterisation of a structural isomorphism TCD which I further develop and defend subsequently in this work. The characterisation differs somewhat from Palmer's position. However, the point is not that his portrayal is incorrect, but rather that his emphasis is somewhat misplaced. The crucial point about structural isomorphism, as a TCD, is that the representation bearers of the representing system are only contentful at all because the value-structure in the representing system mirrors the value-structure in the represented system. When neurophysiological structures or states are organised into a physically structured system that aligns itself with a physically structured system of values of an external feature, the former accrues the capacity to *represent* the latter (the neurophysiological states have the external values as their contents). The isomorphism is instituted, put in place, internally (by the brain), secured through a neurophysiological system exhibiting a certain structure along variations in a neurophysiological variable. At bottom, this is representation *by* magnitude relations rather than representation *of* relations. That is, it is *because* magnitude relations between values in the representing system correspond to magnitude relations between values in the represented system that representation is possible. It is from this that the capacity for the representation *of* relations derives. Palmer's examples present a rather artificial case of a small, fixed set of "known" objects, where it's possible to answer questions concerning, for example, the relative height of any pair of objects in the represented world by appeal to the relative length of lines in the representing world (provided, of course, that the mapping function is known). Consider, however, that what we really require is a system that not so much could "answer questions" about relative height of any pair of "known" objects, but rather one which *represents* height. What is required is a representing system where possible (tokenable) "states" possess values which correspond to the possible values of a representable dimension. This is the more cognitively natural case, and the foundational type of representational capacity. And it is this which constitutes the characterisation of the structural isomorphism TCD being developed in this work.

To further clarify the position on structural isomorphism I've adopted, let me return briefly to Von Eckardt's interpretation. Von Eckardt, recall, sees

Palmer's structural isomorphism theory as one where an isomorphism obtains between a set of relations which constitutes the representation bearer and a set of relations which constitutes the representational object. For instance, the set of relations among the separate lines of varying length in Palmer's representing world example, according to Von Eckardt, constitutes the representation bearer. That is, it is the distinct lines and the height relations between them which makes-up an individual representation bearer. Now, I suggested above that a structural isomorphism TCD need not be viewed this way. And I have been subsequently concerned with putting forward a theory which, like FRS, holds that the isomorphism is between a set of relations which holds *among* a system of representation bearers and a set of relations which holds *among* representational objects (the latter being values of a rather specific represented dimension).

However, even considered restrictively as an interpretation of Palmer, Von Eckardt's portrayal seems somewhat incorrect. Each of the lines in Palmer's representing world differs in length reflecting a different value of the represented dimension. Each of the lines, that is, has a determinate, and different, semantic value or content. This being the case, each line operates as a representation bearer in its own right. The different lines are better viewed, then, as *distinct* representation bearers within a system of such bearers.[77] It is true, as per Cummins' (1996, 96) remark favourably reported above, that each item in the representing system, each representation bearer, is only contentful at all because of the isomorphism between the *whole system* of representation bearers and the represented dimension. This is entirely consistent, however, with the view that items or "states" in the representing system which admit of a determinate content value are operating as distinct representation bearers. So, as suggested, the systemic version of structural isomorphism being developed here claims that sets of relations in the representing and represented systems hold among representation bearers and representational objects, rather than constituting a representation bearer and a

[77] It might be pointed out that the need to have a group of separate lines presented simultaneously is the result of conveying the example in a printed medium. If, instead, we thought of a system capable of going into different states – that is, displaying lines of different lengths – at different times, then the temptation to think of the entire set of states as one representation bearer would be removed. This is certainly correct, and, indeed, a better way of thinking of a system which operates with a structural isomorphism TCD. However, this does not change the fact that Von Eckardt's characterisation of Palmer's theory commits her to the view that the entire set of states constitutes the one representation bearer.

representational object. Indeed, on such a theory, *any* distinct value that can be exhibited by the relevant variable of a representing system is a distinct representation bearer with a distinct content. That is, the representing system consists of all the *possible* values of a certain variable, all the possible "states" the system can go into, or token. The same characterisation applies, *mutatis mutandis*, to the represented system. And the isomorphism obtains between the magnitude relations among the possible values of the representing and represented systems.

2. Structural Isomorphism and Microcontent

What has emerged in the previous section is that the structural isomorphism TCD I've formulated operates with a conception of mental content which I propose to call "microcontent". That is, a representing system represents a rather *fine-grained* feature, or variable, of the represented world. As suggested, any possible distinct value exhibited by the representational substrate instantiates a distinct representation bearer; each bearer's content is a differing value of a certain represented feature. Any complex representational device – including, of course, cognitive agents such as ourselves – will possess a very large number of such representing systems. Content will be fine-grained to the extent that each of the very many representing systems represents a specific variable, and within each system there are a number of possible values that can be exhibited by representation bearers. Over the course of the next two sections, I will describe and defend the microcontent conception. In this section, I begin by fleshing-out the idea of microcontent with particular reference to certain views expressed by Dretske. I then outline some neuroscientific and introspective evidence for the microcontent perspective. Finally, I touch upon the relationship between the microcontent proposal and the selectivity of mental representation.

2.1. Representing and Represented Systems

The microcontent conception departs rather significantly from the standard, rather coarse-grained understanding of mental content. On the latter, the talk is of "objects", literally, or "properties" (for example, the property of being a horse), as the items represented by mental states. Of the more prominent contemporary philosophers of mind, the views of Dretske perhaps come closest to the microcontent conception. Using a non-mental example (which will not effect the point here), Dretske defines his notion of functional meaning (discussed in Chp. Two, subsec. 2.3. above) as follows:

> ...*d*'s being *G* means that *w* is *F* = *d*'s function is to indicate the condition of *w*, and the way it performs this function is, in part, by indicating that *w* is *F* by its (*d*'s) being *G* (1986/1990, 133).

I take it that Dretske's remark here could be interpreted in a manner compatible with the representing and represented *system* conception. *G* is the value taken on by a representation bearer (*d*) in the representing system. *F* is the value of a feature (*w*) in the represented system. Of course, the conception recommended above, in contrast with both Von Eckardt's position and that adopted by Dretske here, is that the representing system consists of a *number* of representation bearers, rather than the one representation bearer taking on different values.

In a later work, Dretske (1995, 2-22) also puts forward, more explicitly this time, the systemic conception:

> A state may derive its indicator function – and, hence, its representational status – from the system of which it is a state. Call these *systemic* indicator functions (functions$_s$) and the representations they give rise to systemic representations (= representations$_s$). If a system (e.g., a thermometer) is supposed to provide information about temperature, and *B* is the state (e.g., mercury at such-and-such level) that is supposed to carry the information that the temperature is, say, 32 degrees, then *B* has the systemic function (function$_s$) of indicating a temperature of 32 degrees. State *B* therefore represents the temperature as being 32 degrees (1995, 12).

Such a view, very much *modulo* the talk of indicator functions, is in line with the representing and represented system conception put forward above. Here we have an example of a representing system (comprised of a number of states, or better, representation bearers) which is dedicated to the representation of a particular external variable (in this case, temperature). Each representation bearer exhibits a value which stands-in for a value (out of a range of values) possessed by the represented feature (the represented system). Dretske, in my view, is right to see representation bearers, and representational objects (values of representable features), as belonging to more or less narrowly circumscribed systems. Moreover, he is right to suggest that each representation bearer within a representing system in some fashion inherits the content it possesses by belonging to a system of representation bearers. On the structural isomorphism TCD, however, the way to unpack such views is to hold that each representation bearer has the content it does because of the objective relation (isomorphism, in this instance) that obtains between the representing *system* to which it belongs and the represented system. To reiterate an important point made in the previous section, it's not that a system represents a variable because each member of the system represents a value of that variable; or (to put it in the form of Dretske's TCD) each member of the system has the function of representing a particular value of a variable. Rather, the objective relation holds between representing and represented systems, not "severally", between each individual member therein. Each representation bearer has the particular content it does because of its "location" in the system. And the objective relation between representing and represented systems is a necessary (but *pace* Dretske, not sufficient) condition for an item's representational status (see Chp. One, sec. 2. above).[78]

Shifting from the realm of artefacts to that of mental representation, Dretske suggests that rather than describing a given sense modality as having the function of representing a certain environmental feature, it would be more accurate to characterise a modality as having "multiple indicator functions." The reason is that a given sense modality can represent different aspects of the represented world. For example, "audition provides information, not only about pitch and intensity, but about timbre and (binaurally) the directional properties of sound" (Dretske 1995,

---

[78] By the way, Dretske's willingness to recognise and broach the issues he (at times) does, somewhat blurs my nice clean taxonomy of "connectedness" and "directedness" conceptions of intentionality (Chp. Three, sec. 1. above). I've portrayed Dretske as a representative of the "connectedness" interpretation, but here we see him moving toward a more fine-grained account of mental content, something more in keeping with the "directedness" interpretation. In fact, some of my earlier discussion already had Dretske as the least dogmatic member of the "connectedness" club. His claim that mental representation bearers have a sense, as well as a referent, and his talk of "purported" representational objects (see, for example, 1995, 32), suggests that Dretske has at least recognised that mental representation bearers represent objects *as being a certain way*, (what I've been calling the "representational-object-as-represented conception of content"). However, the "directedness" and "connectedness" locutions are meant to describe clusters of characteristics or typical factors, not views on or conceptions of certain phenomena that must be all held out of necessity. There's no skin off my nose if certain characteristics of the "connectedness" conception are questioned "from within"; troublemakers are encouraged, and defectors welcome. That said, whether someone like Dretske – someone who advocates the kind of orthodox TCD he does – can make for a *viable* and *robust* notion of, for example, the representational-object-as-represented (and concomitantly, the selectivity of representation), is something of which I am skeptical (for reasons broached in Chp. Three, subsec. 1.2. above).

21-22). Again *modulo* the talk of indicator functions, Dretske's view is in general accordance with the microcontent conception. However, in contrast to Dretske, rather than conceiving of a given sense modality as a representing system representing different aspects of the represented world, I prefer to characterise a given sense modality as comprised of a number of representing systems, and each aspect or variable represented as a distinct represented system. This is a more fine-grained account of representing and represented systems than that employed by Dretske.

2.2. Microcontent: Some Evidence

There is, in fact, considerable neuroscientific evidence in favour of the microcontent conception of mental content (see, for useful introductions, Hubel 1988, and Zeki 1988). It is well known that the brain exhibits anatomical specialisation in regard to the different sense modalities. Of greater relevance to the understanding of "representing system" recommended here is the existence of types of neurons specialised for the representation of specific features within the one sense modality. For example, the evidence points to the visual cortex possessing distinct neural systems for the representation of colour, shape, motion, and depth. Indeed, the evidence suggests functional specialisation of an even finer grain. For example, it appears that certain types of neurons are tuned – that is, respond predominantly – to certain bandwidths of light, while others respond to broadband light. Certain types of neurons are specialised for edge-detection, and others for orientation. And some kinds of neurons respond better to one direction of movement than another.

Examination of the contents of one's conscious experience also reveals support for the microcontent conception. Consider any rich visual experience, such as your current visual experience as you look up from these pages for a moment. Casual inspection discloses that our visual phenomenal field is composed of a number of phenomenal elements: colours, motions, surfaces, edges, and shapes, all at various depths. A visual experience, then, is a composite of a number of phenomenal elements. The overall quality of the complex experience is given by the simultaneous presentation of the contents of its constituent elements (see O'Brien and Opie 1998, 387-388; 1999a, 140-141).[79]

The strategy of appealing to introspective examination of phenomenal consciousness as support for the microcontent conception (and, indeed, the sketch of the nature of representational experience provided above) will likely strike many as odd, even counterproductive or straight-out mistaken. According to the above, our conscious experience at any one moment is a complex, multifaceted amalgam comprised of a range of (in some sense) independent phenomenal elements, which each contribute their individual microcontents. But it has long been orthodox in philosophy of mind to unquestioningly accept that consciousness is *unified* or *singular*, and such a view may well still be held by the majority. According to Paul Churchland (1995, 214), for example, one of the most salient features of consciousness is its unity.

The traditional doctrine of the unity or indivisibility of consciousness, however, has very recently come under increasing pressure, and a growing body of evidence and argument suggests that it can no longer be sustained. Two of the more well-developed recently devised theories of consciousness – Dennett's (1991b) "multiple drafts" proposal, and O'Brien and Opie's (1999a, 1999b) connectionist account – both dispute (despite the significant gulf between their respective theories) the legitimacy of the orthodox conception of consciousness. Both Dennett and O'Brien and Opie put forward *"multi-track"* theories of consciousness (the term is due to the latter theorists). On such a view, there is no single consciousness-making mechanism or process located in the brain. Rather, phenomenal experience is "distributed"; mechanisms responsible for generating consciousness are to be

---

[79] It might be supposed that such an account of the contents of phenomenal consciousness walks straight into the clutches of the so-called "binding problem". Since the above account supposes that the activity of various types of representing systems is not passed to a single area of the brain where it is combined, there is, objectors will argue, a question as to how distinct types of representational elements coalesce to yield our familiar perception of objects. However, O'Brien and Opie (1999a, 143-144) argue that the binding problem admits of a relatively straightforward solution. On their view, it is the mere *simultaneity* of the various elements produced by the specialist bearers which results in the required binding. That is, there is no need for the activity of the bearers to be passed to the one location when simple co-occurrence will do the job. Certain neuroscientists have also developed a response along such lines (see, for example, Edelman 1987).

found across spatially distinct areas of the brain.[80] Each consciousness-making mechanism *individually* contributes a distinct phenomenal element. The multifarious nature of our overall phenomenal field consists in the simultaneous presentation of various phenomenal elements, resulting from the activity of a whole host of spatially distributed consciousness-making mechanisms. So, consciousness is not singular or unified in its *production*. And if there's no single consciousness-making mechanism, the idea that consciousness itself is singular looks to be undermined.

O'Brien and Opie (1998) take the debate one step further. They argue that the project of seeking to locate a single part of the brain as the source of phenomenal consciousness is wholly, and "conceptually", misguided. On their view, such a futile search is brought about by a mistaken conception of the sense in which consciousness is single or unified. I said above that consciousness, on a multi-track approach, is not unified in its production. Surely however, many will object, consciousness is unified in its *presentation*. O'Brien and Opie submit such a claim to critical scrutiny, urging caution along two dimensions. For one thing, they argue that the traditional view of the singularity of consciousness is a result of conflating consciousness with the self, the sense of *subject unity* (1998, 393-394). Conscious experiences do not just happen, they occur to an individual subject. Every one of your diverse conscious experiences comes marked with your insignia: they "belong to" you, or better perhaps, in some sense "constitute" your self. How, precisely, the brain manages the considerable achievement of constructing this sense of subject unity is a matter of which there is increasing speculation.[81] Second, O'Brien and Opie acknowledge that there is a sense in which it is legitimate to conceive of consciousness as unified, but it is not that of singularity or "oneness". Rather, consciousness is unified in its *"harmony"* or *"coherence"* (1998, 386-387).

[80] The central argumentative theme of the "negative strand" of Dennett's treatise on consciousness, *Consciousness Explained*, concerns the demolition of what he thinks is the traditional perspective on consciousness, namely, the "Cartesian theatre" model. On such a view, there is a single neurophysiological structure or system where the contents of consciousness band together to put on a show for the benefit of the "mind's eye". As an alternative, Dennett develops the view that consciousness is a global or systemic property, a feature of an organism composed of the monopolising effects of content fixations (Dennett calls them "microtakings") distributed across the brain on cognitive activity and behaviour. O'Brien and Opie (1999a, 1999b) likewise reject the view that there is a single neurophysiological structure responsible for the generation of consciousness. On their connectionist proposal, the contents of consciousness are identified with stable patterns of activation across PDP-style networks in the brain. Such stable, PDP representation bearers are found, they claim, across spatially distinct areas of neural hardware.

[81] For a couple of suggestions see Dennett (1991b, Chp. 13) and O'Brien and Opie (1998).

Instantaneous conscious experience involves presentations from various sensory modalities which coincide in their representational content: "we *see* our bodily parts in positions we *feel* them; we *hear* sounds emanating from objects in the direction we *see* them; we *taste* the food that we can *feel* in our mouths; and so on." Such coherence goes right down, I would claim, to the microcontent level: we see the *texture* of an object ending at the object's *edges*; we hear the *timbre* of a voice emanating from the same source as its *pitch*.

So, consciousness is unified neither in its production nor presentation, at least not, as regards the latter, in the way the traditional doctrine of the unity of consciousness would assert. And I can legitimately appeal to introspective examination of representational experience as support for the microcontent conception. Such introspective evidence can, I've shown, be buttressed by certain theoretical considerations.[82]

The neurophysiological and introspective evidence for the microcontent conception coincide strikingly in cases of patients suffering from various kinds of neurophysiological trauma. Damage to particular neurophysiological sites leads to the disappearance of particular aspects of phenomenal experience, without detrimental effects upon all the other elements. Zeki (1993), for example, reports the case of a woman who lost the ability to visually experience one element of the phenomenal field, namely, *motion*. The woman suffered from a lesion to a part of the brain outside the primary visual area, a part the microcontent conception would propose as the motion representing system (for visual perception). Zeki relates the unfortunate consequences for even simple, everyday tasks:

> She had difficulty, for example, in pouring tea or coffee into a cup because the fluid appeared to be frozen, like a glacier. In addition, she could not stop pouring at the right time since she was unable to perceive the movement in the cup (or a pot) when the fluid rose. The patient also complained of difficulties in following a dialogue because she could not see the movement of…the mouth of the speaker. (Quoted in Zeki 1993, 82).

[82] It is pertinent to point out that O'Brien and Opie (1999b), in their reply to commentators, wed their connectionist theory of consciousness to the structural isomorphism TCD. Indeed, they rely on the structural isomorphism theory to counter certain objections raised against their account of consciousness (see, in particular, 1999b, 180, 182). I'll examine structural isomorphism in the light of connectionism in Chapter Six.

Tellingly, Zeki points out that while the woman lost the ability to experience movement in the domain of visual perception, she had some conscious perception of movement "elicited by auditory or tactile stimulation" (1993, 82). The woman's injury was restricted to just one isolated area of the brain, and she did not seem to suffer from any other cognitive deficits, just as the microcontent hypothesis (and multi-track theories of consciousness) would expect. Another illuminating case comes from patients recovering from visual cortex damage. The phenomenology of such patients reveals that the various phenomenal elements *separately* reassert their contribution to the overall phenomenal field. Consider the following evidence reported by Smythies.

> At first the patient will see pure motion (usually rotary) without any form or colour. Then brightness perception returns as a pure Ganzfeld – a uniform brightness covering the whole visual field. When colours develop they do so in the form of 'space' or 'film' colours not attached to objects. The latter develop as fragments which join together and eventually the colours enter their objects to complete the construction of the phenomenal object (1994, 313).

Smythies suggests, quite plausibly, that the systems responsible for the distinguishable phenomenal elements differ in their rates of recovery.

2.3. Selectivity

I want to finish this section by remarking on the connection between the microcontent perspective and the selectivity of mental representation. Selectivity, as indicated above (Chp. Three, subsecs. 1.2., 1.3.), refers to the idea that mental states don't represent objects *tout court*; rather, they represent objects as being a certain way. Mental states represent certain features of objects and are silent about others. The represented features can be rather coarsely or rather finely individuated. And "within" the representation of a certain feature (certain variations along the values it exhibits), mental representation bearers operate at a certain degree of resolution or grain. Selectivity is bound up with the accuracy – the *variable accuracy* – of representation. Different kinds of representational creatures, and different kinds of representing systems, are differentially accurate. Complex representational creatures such as ourselves have representational capabilities that are – both overall, and in some (but not all) cases, in terms of particular representing systems – more "expressive" than those of simpler cognitive creatures.

There are, then, three more or less distinct "elements" of selectivity, and thereby the accuracy of representation: (i) the number of representing systems implicated (how many features or variables of the representable world are the preserve of a dedicated representing system); (ii) how finely individuated each representing system is; and (iii) the resolution along variations of the representing property (the number of representation bearers comprising a representing system). A cognitive *creature* will possess a more accurate – more expressive – representational capability than another in terms of (i)-(iii). That is, it represents a great number of features of the representable world, the features it represents are finely individuated ones, and its individual representing systems operate with a high degree of resolution. Representing *systems* have a high degree of accuracy in terms of (ii) and (iii); that is, narrowly circumscribed represented features, and a large number of distinct representing bearers permitting fine discriminations along values of represented variables.

Inter-species differences in representational capability are an obvious case of differentially accurate cognitive creatures. It's a fact that different kinds of creatures exhibit differences in terms of the number of features of the world they are able to represent. For example, most kinds of organisms lack the ability to represent colour, a representational capacity that the vast majority of us enjoy. The neuroscientific evidence for the microcontent conception discussed above reveals that differences in the number of features represented also holds in an intra-species fashion, in the unfortunate cases of particular representational capacities (such as motion, or colour) being lost while others remain intact. Differences in representational accuracy among individuals of our own species more typically manifest themselves in terms of (ii) and (iii) above, a clear case in point being expertise. I take it that wine judges, for example, have more finely individuated (or, at least, more clearly delineated) dimensions to their olfactory and gustatory representational capacities, and make more fine-grained discriminations along such dimensions, than those of us who "don't know a lot about wine, but know what we like." The same type of story is to be told, I presume, about differences between certain elite athletes and sportspeople and the rest of the population as regards proprioceptive representational capacities. There are also interesting and marked inter-species differences in representational accuracy which look to be largely a product of differences in the resolution of the relevant representing systems (that is, (iii) above). It is commonplace knowledge that our canine companions, for example, have significantly more fine-grained olfactory discriminative powers than we possess (see, for example, Churchland 1995, 26-27).[83]

Once the undeniable reality of variable accuracy in cognitive representation is recognised, attention to the selectivity of representation is demanded. In the literature, however, selectivity is rarely acknowledged, let alone being the subject of explicit and detailed discussion. I've claimed, in contrast, that selectivity is a fundamental characteristic of mental representation. And I've indicated that I seek to explain content determination in a way that respects such a salient characteristic. I take it that it's not difficult to see how structural isomorphism, wedded to the microcontent proposal, incorporates selectivity. Indeed, the microcontent conception handles selectivity *by definition.* I defined the microcontent perspective in such a way that it cannot help but accord with the selectivity of mental representation. Microcontent posits a distinct, proprietary representing system for each distinct variable a cognitive creature represents, claims that variables represented are specific or finely individuated ones, and proposes that each representing system represents a range of possible values exhibited by the represented system at a certain degree of resolution. Differential selectivity – variable accuracy – is accounted for in terms of differences in one or more of the three elements.

Microcontent, then, unashamedly builds-in selectivity. This is *just* what we require. Characteristics of mental representation should *straightforwardly* follow from the conception of mental content and the theory of mental content determination. And it is the essence of structural isomorphism/microcontent to be selective. I'm reminded here of a point made by Fodor and his colleagues when they first drew attention to what they claimed to be a fundamental characteristic of cognition, namely, the systematicity of thought (Fodor and Pylyshyn 1988; Fodor

[83] While a highly developed representational capacity is an important component of expertise, it is not, of course, the whole story. Expertise is also a product of the processing of representation bearers. A perceptual experience of a work of art had by an art critic, I think it's plausible to assume, has richer epistemic liasons than the perceptual experience had by an everyday member of the population.

and McLaughlin 1990).[84] Fodor and his fellow classicists deploy systematicity against the connectionist account of cognitive architecture, claiming that it's not enough to show that a cognitive architecture is *compatible* with systematicity (that thought *can* be systematic), rather it needs to be shown that a cognitive architecture *entails* systematicity (why thought *must* be systematic). And this, they think, is something classicism, but not connectionism, does. I would make a similar, though perhaps slightly weaker, kind of claim about selectivity. Proponents of one or other of the orthodox TsCD *may* be able to show that selectivity is possible on their favoured account of content determination (though I've indicated certain reservations; see Chp. Three, subsec. 1.2.). But it's one thing to show that an orthodox TCD can be made to fit with selectivity. It's another to have selectivity naturally and straightforwardly follow from the TCD; that selectivity is embodied in the TCD by definition.[85]

3. The Determinacy of Content (Too Many Isomorphisms?)

I concluded the previous section by remarking on structural isomorphism's explanation of selectivity. That explanation, however, is threatened if the claim that a representing system represents a *distinct* representable variable cannot be secured. Indeed, such a worry threatens more than just the account of selectivity. It constitutes a potentially damaging objection which can be leveled against the very heart of the structural isomorphism TCD. The claim is that content determination cannot be a matter of structural isomorphism because structural isomorphism is too promiscuous: any representing system will be structurally isomorphic to a large number of represented systems. That is, if it is plausible to claim that there obtains a structural isomorphism between any given representing system and a represented

---

[84] 'Systematicity' is the idea that thoughts come in semantically related clusters. If a cognitive creature has the capacity to think a certain thought (such as, "John loves the girl"), then it also has (indeed, must have) the capacity to think a range of semantically related thoughts (such as, "the girl loves John"). We don't come across cognitive creatures with punctate minds; cognitive creatures, that is, whose cognitive capabilities consist, for example, "of the ability to think seventy-four unrelated thoughts" (Fodor and Pyslysyn 1988, 40).

[85] The reason why my claim here is perhaps slightly weaker than the claim about systematicity is that Fodor and his colleagues assert that systematicity is an *essential* feature of thought, a matter of *psychological law* (see Fodor and Pylyshyn 1988, 40; Fodor and McLaughlin 1990, 184-185, 202). I'm not sure that I'd want to claim that selectivity is an essential feature of mental representation (though, as a matter of contingent fact, I think that the representational capabilities of cognitive creatures on this planet exhibit selectivity). It's possible to imagine, I suppose, cognitive creatures whose mental lives don't manifest selectivity. And perhaps certain of our higher-order, language-infested thoughts are not selective to any significant degree.

system, then it's surely the case that a representing system will exhibit a structural isomorphism to a *range* of representable variables. And this means that representational content is indeterminate. The content of a representation bearer belonging to any given representing system is not the value of a certain representable feature *f*, but the values of representable feature *f*, representable feature *g*, representable feature *h*, and so on.[86] In this section I will put forward three lines of response to this objection. I do want to recommend one of these types of response over the others. For those who don't share my view, however, this section should at least serve to show that the advocate of structural isomorphism is not without resources when it comes to dealing with the objection at hand.

To begin with, notice that the objection is not as simple as it has been portrayed so far; what's more, the intended conclusion (that the structural isomorphism TCD is not viable) doesn't straightforwardly follow from the assertion that a representing system will exhibit a structural isomorphism to many represented systems. Let me flesh-out the objection. The foundational claim is that any system of representation bearers will be structurally isomorphic to a number of different aspects or features of the represented world. The second claim, suppressed in the above presentation, is that there is no principled means by which a representing system can be said to represent one, rather than a whole host, of represented systems to which it bears a structural isomorphism. The third claim, also implicit above, is that there is something unpalatable about a representing system exhibiting a structural isomorphism to a number of different represented systems. The idea here is that Fodor, Dretske, Searle, and the like are right: mental states are originally and determinately contentful. On the basis of these three claims, the intended conclusion is, of course, that structural isomorphism is not a viable contender as an account of mental content determination. In what follows I will consider three types of response to this "indeterminacy objection", one for each of the above three claims.

[86] This objection to the structural isomorphism TCD was first put to me by Gerard O'Brien, who encountered it at a presentation of his 1998 paper (personal communication). I'm not suggesting, however, that O'Brien thinks the objection cannot be met. Turning to the literature, Cummins (1996, 102-103) clearly seems to think that structural isomorphism renders content indeterminate.

3.1. Two Replies

I begin with the third claim, namely that there is something undesirable about the sort of indeterminacy alleged to arise from the structural isomorphism TCD. One line of response would be to agree that mental states are originally contentful (the structural isomorphism TCD, because of its postulation of an objective relation between representing and represented systems, certainly doesn't suggest otherwise), and accept, moreover, that the structural isomorphism TCD has the sort of indeterminacy described above. The line taken here, however, to put it in broad terms, is that this is the kind of indeterminacy we can live with. That is, this type of reply concedes that any system of representation bearers will be isomorphic to a number of different represented systems, but denies that the sort of indeterminacy here is a problem.

I think there are two possible versions of this sort of response. The first – call it the "interpretant" version (see Chp. One, subsec. 2.1.) – argues that any mental representation bearer belonging to a given representing system has a *determinate effect* on the cognitive agent. Specifically, the tokening of a representation bearer modifies the agent's behavioural dispositions. The particular modification made depends on (let's assume) the particular physical profile or "shape" of the tokened representation bearer. And this is as determinate as you like. The representation bearer has a determinate physical profile, and thereby effects a determinate modification to the cognitive subject's behavioural dispositions. As outside observers, we might not be able to say (for certain) what represented feature the representation bearer is standing-in for. But it's not the case that the representation bearer is indeterminate in its *value*, since its value is given by its "place" in a system of representation bearers. And there is a fact of the matter as to what this value is. So, according to this version, content is not indeterminate for the interpretant of a mental representation bearer.

The other variant on the first line of response can be termed the "application" version (after Cummins 1996; see Chp. Three, subsec. 1.4. above). The idea here is that the representation bearer will be applied to, or used to represent, a particular representable feature. That is, a representation bearer is not simultaneously applied to a number of features of the represented world, but rather to one feature. Like the first version, this version says that what a representation bearer represents is not, "in practice", indeterminate for the cognitive system harbouring the representation bearer. A representation bearer is not indeterminate as regards its effects, the "use" to which it is put.[87]

In my view, this first line of response is unsatisfactory. In some way, both versions attempt to say that there is something determinate about the role or use of a representation bearer. But as I've been at pains to point out above (Chps. Two and Three), what a representation bearer represents must be determined independently of its use. It is difficult to contemplate collapsing the two again here (even in this slightly different way). Moreover, surely an advocate of original intentionality is going to want more than what the first reply provides. A defender of original intentionality will claim that a representing system has *a* represented system (for one thing, it certainly seems that way to introspection), and this is something a TCD must secure. Let me then move on to the second type of response, one which attempts to make good such a claim.

The second type of reply is directed at the second claim of the indeterminacy objection, namely, if a representing system exhibits a structural isomorphism to a range of representable features, there is no way of showing that it has only one of these features as its represented system. The response is that this is simply false. In fact, the line of thinking goes, there are a number of ways a represented system might be determined: any one of the orthodox TsCD can be viewed as a way of securing the feature a representing system represents. In effect, the standard intuitions about content determination are best viewed as possible accounts of how to determine a represented system. The idea here, then, is that the structural isomorphism TCD be supplemented with some additional equipment. While structural isomorphism is an account of the values a system of representation bearers stand-in for, some other story is to be told about what feature the system of representation bearers represents. And, as suggested, any one of the orthodox TsCD can be wheeled-in to tell this story.

So, for example, one could argue that the feature a representing system represents is the one presented during the learning period, that feature on which the system was trained. This is a Dretske-like manoeuvre (see Chp. Two, subsec. 2.2.) applied to the structural isomorphism framework. Alternatively, one might opt for a

[87] Cummins (1996, 121) would appear to endorse this kind of reply to the indeterminacy worries of structural isomorphism.

more standard causal account. That is, there is a causal connection between a representing system and its represented system. The represented feature is that which brings about activity in the representing system, that which results in the tokenings of representation bearers. Or, one might go teleological. The represented system is that which the representing system has the function of representing, that which it was selected for.

Again, however, I find this type of reply unsatisfactory. For one thing, there is no reason to think that the problems encountered by the orthodox TsCD in their usual role will not resurface here (these difficulties were discussed above in Chp. Two, sec. 2., and Chp. Three, subsec. 1.2.). Of particular relevance, the orthodox theories, as has been shown, have indeterminacy worries of their own. That is, aside from concerns about their internal coherence and any other reservations one might have, each of the orthodox TsCD is plagued with indeterminacy troubles which threaten the viability of the TCD itself, as well as undermining the standard proposals for accommodating misrepresentation. Second, it is somewhat *ad hoc* to tack-on another theory to the structural isomorphism TCD in order to counter an objection. Surely it would be preferable to counter the objection at hand using only the resources of the structural isomorphism TCD. The third reply says that the threatened indeterminacy can be avoided from within the structural isomorphism framework itself.

3.2. The Structural Isomorphism Reply

In my view, the best option for the structural isomorphism theorist is to take a "head-on" approach to the indeterminacy objection. On this third line of response, the very plausibility of the foundational claim is questioned. In the (admittedly sparse) presentations of the indeterminacy objection I've encountered, there seems to be a straightforward acceptance of the idea that a representing system will manifest structural isomorphisms to many variables of the representable world. But I'm not aware of any considerations offered in support of such an idea. The claim, as far as I can tell, is mere supposition. And in the absence of any supporting considerations, the way is open for the advocate of the structural isomorphism TCD to simply disagree with the first claim of the indeterminacy objection. However, the dispute then rests with what is basically a clash of intuitions, which is not particularly satisfying. In what remains of this section, I will put forward a consideration which suggests that the intuition grounding the indeterminacy objection is misguided. I will touch on two further considerations to the same end in following chapters.

It's worthwhile examining why someone might be moved to propose the "many represented systems" claim in the first place. Why does it initially seem so plausible? The *prima facie* plausibility of the claim, I think, derives from an unimaginative conception of a representing system (in particular, a neural representing system), and thereby an impoverished conception of the structure among representation bearers embodied by a representing system. Allow me to introduce some terminology, and call the structure exhibited by a representing system an *"analogue"*. My claim, then, is that the intuitive tug of the indeterminacy objection's foundational premise rests with an overly simplistic understanding of an analogue. Such a conception is abetted by thinking solely in terms of very simple representational devices, those which subserve very primitive representational capabilities. And it's also a product of thinking of a representing system in isolation, as opposed to a representing system as a constituent of a sophisticated organism's cognitive capabilities (as opposed, that is, to a representing system's contribution to the overall cognitive and behavioural repertoire of a complete cognitive agent).

The mistakenly crude conception of an analogue, then, is due to an overly narrow focus. Once we shift our focus to reasonably complex cognitive systems, systems capable of subserving a significant degree of behavioural flexibility, the conception of representing systems as rather crude structures becomes increasingly difficult to sustain. The analogues possessed by any half-way decent cognitive creature will be highly elaborate structures. A *multipurpose* visual system, for example, one which could be deployed in performing a multitude of tasks, would have a highly complex internal constitution. The system as a whole would need to be made-up of a number of narrowly circumscribed, proprietary representing systems, *each* of which represented an *aspect* of the environment in a fine-grained fashion. It would need to be able to discern, that is, those different variables of the visually representable world, and subtle variations in the value of each representable variable. Reflecting on cognitive creatures with a sophisticated behavioural repertoire, creatures capable of interacting with a diverse environment in a largely appropriate manner, suggests that we need to think of representing systems as being very *complex, intricate, and idiosyncratic*. And as we move to a conception of each representing system as embodying its own highly complex and particular organisation, it becomes harder to maintain that each such representing system will exhibit a structural isomorphism to a number of features of the representable world. Each individual structural isomorphism implicated in mental representation is going to involve a very complicated structural correspondence.

Interestingly, the starting-point of the above line of reply parallels Dennett's (1987, 29-33) response to indeterminacy worries brought against his own TCD. The threat of indeterminacy is particularly salient for Dennett, since he rejects the idea that content is fixed by some kind of objective relation between representation bearers and representational objects (see Chp. One, subsec. 1.2.). On Dennett's account, to reiterate, if a system's behaviour can be reliably predicted by the attribution of a certain set of contentful states, then that is *"all there is"* to the system possessing those contentful states. There is nothing deeper which serves to underwrite representational content. Dennett readily accepts a straightforward consequence of his position: it may turn out that rival predictive schemes are equally successful, and thereby the content of the subject's representational states is ultimately indeterminate. Stated thus in its bare form, and in isolation, Dennett's account of content determination looks to have radical content indeterminacy follow as the *normal* case. But Dennett, to my mind successfully, takes the bite out such indeterminacy worries. While he accepts that indeterminacy is always in principle possible, in reality – in the cognitively natural case – it is rarely at issue. According to Dennett, once a system has rich sensory attachments to its environment – once a relatively complicated cognitive system is up and running in the world – a unique predictive scheme is, in practice, all but demanded. He writes:

> ...the class of indistinguishably satisfactory models of the formal system embodied in its internal states gets smaller and smaller as we add such complexities [of sensory attachments]; the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated (1987, 30-31).

According to Dennett, then, once we shift our focus from the TCD in isolation to a full-fledged cognitive system in the wild, the threat of indeterminacy is practically negligible.

I've applied the same type of consideration to the structural isomorphism TCD. Stated in its bare, theoretical form, structural isomorphism likewise looks to be vulnerable to content indeterminacy, via the supposed unwanted proliferation of isomorphisms. The telling move is to conceive of structural isomorphism as situated in a behaviourally versatile cognitive agent. Once structural isomorphism is given something to do – namely, underwrite a cognitive agent's behavioural flexibility (and this is the whole point of proposing representationalism in the first place) – the charge that any representing system will embody an analogue of a range of disparate representable variables loses its intuitive force. The indeterminacy objection rests on an inadequate understanding of the nature of a representing system. Any given system of representation bearers must be understood as a very intricate and idiosyncratic structure. Structural isomorphisms between neural representing systems and environmental variables will be highly complex (indeed, as will become apparent in Chapter Six, so abstract that though they look to be there, they are difficult to pin down precisely). Isomorphisms are just not, as the indeterminacy objection would have it, ubiquitous.

To conclude my favoured reply to the indeterminacy objection I wish to offer a few – admittedly somewhat speculative – remarks about the kind of complexity exhibited by the representing systems of the structural isomorphism TCD. The remarks are somewhat sketchy, but this is only to be expected in the early days of the development of a new theory. Some illustration of the points to follow will be provided at subsequent stages in this work.

The first remark is that the value-structure of the representing system (and indeed the represented system) will exhibit a certain "grain", or resolution. This means, for one thing, that there will be a certain number of values, and a certain range of values, that the variable of the system can express. In other words, there will be a certain number of "states" the system is capable of going into, with an upper and lower limit on the possible values. For another thing, the grain of a system will be determined by the degree of difference between any pair of values in the range of possible values. Of course, this itself is a factor of the number of possible distinct values the variable can exhibit. A coarse-grained system, then, would be one where only a relatively small number of values were possible, with relatively large differences between each possible value. A fine-grained system would be able to go into a relatively large number of states, each state only slightly different to those of neighbouring values. Such a representing system would present a more precise representation of the represented feature.

The grain of a representing system may itself possess a further interesting characteristic: it may not be uniform; it may exhibit certain "deformations". In other words, there may not be, throughout the range of possible values, constant differences between them. At certain areas of the range, the values might be densely packed together. That is, a large proportion of the possible states of the system may be collected in a certain area of the range, with very small differences between each of the values. Other areas of the range may be more sparsely populated with possible values, with larger differences between each of the values. Indeed, there are all sorts of possibilities regarding the distribution of values throughout the range of possible values. The system might be linear in the sense that the degree of difference between the values changes in a constant function as we move from one end of the range of values to the other. So, for example, the system may be maximally fine-grained at the upper end of the value range, gradually and constantly becoming less fine-grained as we move to the lower end of the value range. The converse is also possible. Alternatively, the system may be most fine-grained at regions near the upper and lower limits of its range. Or the distribution of values might be more or less Gaussian, with most of the possible values located at the centre region of the value range, the density gradually trailing off as we move further from the area of moderate values. Most radically, the system might be non-linear in that there is no systematic change in the differences between values throughout the range of possible values. Indeed, the possibilities in regard to the "landscape" of system grain are endless. Moreover, any representational entity is going to possess a number of representing systems, and there is no reason why each system might not exhibit a different grain.

We now have, then, a rather more complex picture of the structural isomorphism TCD than suggested by the earlier formulations. Palmer's simple examples which I employed above, while useful as introductory illustrations of structural isomorphism, also have the potential to mislead. The structural isomorphisms involved in mental representation will be highly complex. The indeterminacy objection, I suggest, is merely an unexamined intuition. I have considered three ways in which it might be overcome. Ultimately, I've recommended that the *prima facie* plausibility of the objection can be undermined using the resources of the structural isomorphism TCD itself. This, I think, is the right way for the advocate of structural isomorphism to go. The heart of my favoured reply to the charge of too many isomorphisms is now in place, but I do propose to touch upon the objection on a couple of occasions below. There I will seek to buttress my response with further considerations that demonstrate just how complex will be the structural isomorphisms involved in mental representation.

4. Structural Isomorphism and Misrepresentation

In Chapter Two I commenced a line of argument which pursued content determination and the problem of accommodating misrepresentation. I began with the problem of misrepresentation and the orthodox approach. The orthodox approach was further examined in Chapter Three, and subsequently contrasted with the emerging alternative approach set-out in section 2. of that chapter. The route I've traced from misrepresentation through the development and defence of the structural isomorphism TCD culminates in this section. Structural isomorphism provides the kind of TCD required of the alternative framework. Now that the structural isomorphism TCD has been formulated, I intend to rehearse the shape of this proposal for content determination and misrepresentation, and touch upon certain of its advantages (subsec. 4.1.). To complete the investigation into misrepresentation, I then turn to the issue of the determination of targets or intentional objects (subsec. 4.2.).

4.1. Content Determination and Misrepresentation: The Alternative Approach

In the summary that concludes Chapter Three, I put in place a framework for an alternative approach to content determination and misrepresentation. The new framework took shape from a rejection of the assumptions and commitments of the established strategy. The alternative framework does not begin with veridical content, with correct application. Rather, it completely decouples the veridicality of content from the determination of content. The theory of content determination is blind to whether content is true or in error, since content is not of itself veridical or in error; rather, the question of veridicality is an orthogonal issue, one which asks whether a given representation bearer has been correctly or mistakenly applied. The account of misrepresentation utilises resources beyond those implicated in content determination.

The alternative framework is not forced into the problematic and so far unsuccessful tactic of attempting to show how content and application come apart. As such, it gains a crucial extra degree of freedom. It breaks free of the idea that both content determination and misrepresentation can be handled with the one objective representation-bearer-to-representational-object relation, and thereby the strategy of specifying a special subclass of bearer-to-object relation. The new framework has it that content and intentional object (the object to which a bearer's content is applied) are fixed *independently*. And since they are independently determined, they can fail to coincide. The alternative framework drives a wedge between content determination and application right from the outset. This is *just* what's required of a *principled, natural, and straightforward* means of accommodating the capacity for misrepresentation.

What such a solution to the problem of misrepresentation needs, of course, is a use-independent explication of content determination, and an account of intentional object fixation. On the former, content determination must be antecedent to and independent of the use of representation bearers, and this means the orthodox TsCD must be rejected. A use-independent notion of content determination, on the other hand, is what structural isomorphism, and only structural isomorphism, affords. Structural isomorphism, because it is founded upon the relational physical structure embodied by a proprietary representing system, has content determination as a function of the physical properties of representation bearers themselves, not something about the causal relations into which they enter, the use to which they are put. Content determination and application are kept distinct. Structural isomorphism provides the TCD demanded of the alternative framework. All that remains to transform the proposed rival framework into a full-fledged alternative approach is an account of intentional object determination. That is the task I undertake in the following subsection. Before I get to that discussion, however, I have a couple of remarks to make.

The line of argument I've been pursuing has a subtle implication which further mitigates in favour of the alternative approach formulated here. The fact is that misrepresentation, on *any* approach, is, by definition, a deviation or mismatch of content and intentional object. We can think of any proposed solution to the problem of misrepresentation not just as a way of allowing for such a mismatch, but as also tacitly embodying an account of *"why"* misrepresentation occurs. As I've made plain, the orthodox approach begins by conflating content and correct application, and then seeks to show how content and the intentional object can come apart. However, I've argued that none of the particular incarnations of the orthodox approach satisfactorily resolves the problem of misrepresentation. And without a viable account of the "what" of misrepresentation, advocates of the orthodox strategy are left with nothing to say about why misrepresentation is a fact of cognitive life. In contrast, not only does the approach to content determination and misrepresentation I've proposed provide for a natural and straightforward means of admitting misrepresentation, it yields a satisfying account of *"why"* misrepresentation occurs. Content and intentional object can deviate – fail to coincide – because they are *independently* determined. Being a product of two independent forces, it's no surprise that they will sometimes fail to march in step. Indeed, as shown in the discussion of teleosemantics, sometimes it's of positive benefit that the two should be separately determined; misrepresentation is sometimes a very useful strategy in the cut and thrust world of survival with its constraints of time-dependent action and limited cognitive resources. On general evolutionary grounds, however, it's also not surprising that Mother Nature has got content and intentional object to routinely line up in her more behaviourally versatile and successful cognitive creatures.

So the alternative approach, with its crucial extra degree of freedom, provides a natural and appealing account of why misrepresentation is a fact of cognitive life. Of course, a more complete answer to why misrepresentation has occurred would appeal to the intuitively plausible idea that something has "gone wrong" with the cognitive agent. In certain cases of radical misrepresentation – such as the hallucination of a dagger – something has gone very wrong; to find out what, we'd have to investigate the agent's cognitive machinery, including the recent history of its operation. What Macbeth "needs" to represent is an area of "unoccupied" space before his eyes. In his distressed state, Macbeth undergoes a visual experience with the internally-generated content of a dagger, which is then applied to an area of "unoccupied" space. Content and intentional object are independently determined, and deviate strikingly in cases of radical misrepresentation.[88]

The alternative approach can also make sense of less radical cases of misrepresentation, such as mistakenly perceiving a cow as a horse. Indeed, here it has even more to say, along the following lines. Visual experiences are, obviously, standardly elicited by the external world. No one could sensibly deny that states of the external world cause the tokenings of mental representation bearers. What the advocate of structural isomorphism does deny, of course, is that content determination is to be understood in terms of such a world-to-head causal relation. While experiences are standardly elicited by external conditions, their representational content is internally determined. Now, why should the presence of a cow bring about the tokening of a "horse" representation bearer? Well, according to the alternative approach, the presence of a cow in the immediate environment brings about neural activity in the cognitive agent who is cow-confronted. The degree of this activity occasions (in a proprietary representing system) the tokening of a representation bearer with a certain value of a variable as its content.[89] The degree of activity influences what representation bearer is likely to be tokened, but, again, does not determine the bearer's content. The "values" of the two external objects at issue are quite similar (a cow is more like a horse in appearance than it is like, for example, a 1971 Holden Monaro). Consequently, the neural impact occasioned by each kind of object is quite similar. And, of course, it's the claim of the structural isomorphism TCD that the brain's representing systems consist of physically-structured sets of representation bearers, with similarity of representational content reflected in similarity of physical properties. It is, then, no surprise that the presence of a cow should sometimes result in a "horse" representation bearer on the view of representing systems proposed by structural isomorphism. The cognitive impact of being cow-confronted is activation of that "area" of a representing system where representation bearers with contents "cow" and "horse" are closely "located". And, as shown above (Chp. Three, subsec. 1.4.), in the perceptual case it is typically the external tokening object to which a resulting

---

[88] As I'll go on to show in the following subsection, the state of the world is one component of intentional object fixation.

[89] This is a toy example because, of course, such a case – according to the structural isomorphism TCD – would actually implicate a number of representation bearers. Nonetheless, it does serve as a useful illustration.

representation bearer is applied. The cognitive agent mistakenly applies a "horse" representation bearer to the external object which generates its tokening, namely, a cow, thereby resulting in a case of mild misrepresentation. Again, the story here accords with the appealing intuition that something has "gone wrong" within the cognitive agent; in this case, mildly wrong.[90] *Why*, in contrast, should a cow elicit a "horse" representation bearer on an account of content determination which does *not* posit the kind of dedicated, physically-structured systems of neural representation bearers proposed by structural isomorphism? An account of content determination other than structural isomorphism leaves the "why" of mild misrepresentation mysterious.

Finally, I want to highlight how neatly the structural isomorphism TCD handles what Fodor (1990a, 91) calls the "robustness" of content. Robustness is a *desideratum* of a satisfactory TCD, and refers to the idea that any given type of mental representation bearer can be tokened by *all kinds* of conditions (a veritable plethora of environmental conditions, and, potentially, any other kind of mental state) without this factoring into the determination of content. That is, any mental representation bearer's content is insensitive to the heterogeneity of causes of its tokenings. Robustness is pretty obviously a particularly acute problem for causal theories of content determination. As Fodor (1990a, 91) writes: "If there's going to be a causal theory of content, there has to be some way of picking out *semantically relevant* causal relations from all the other kinds of causal relations that the tokens of a [representation bearer type] can enter into." As I've shown at certain points herein, this is by no means a straightforward task, even in the minimally limiting case of solving the disjunction problem (of admitting misrepresentation). The structural isomorphism TCD, in stark contrast to the causal theory, completely bypasses any worries about the robustness of content being violated. Content is fixed without any reference to the causal relations into which mental states enter, both those environmentally-instituted causal relations which result in the tokening of mental states, and those causal transactions mental states conduct with other mental states and behavioural outputs. Robustness is a deep fact of cognitive reality; it pervades a cognitive life. Structural isomorphism *inherently* respects robustness.

---

[90] Something can "go wrong" under *any* condition. Or, no circumstance guarantees correct representation, which is one reason why, as I've claimed, the orthodox approach is misguided.

4.2. The Determination of Intentional Objects

Distinguishing content determination and application, and formulating a use-independent TCD, "breaks the back" of the problem of misrepresentation. This is significant progress, however the alternative approach is only rendered complete with an account of the fixation of targets or intentional objects, the final issue I'll address in regard to misrepresentation. Here I defer largely to Cummins (1996, Chps. 2 and 8). In overview, the determination of "targets" (as Cummins calls them) or "intentional objects" (on my terminology) is a matter of two factors[91]: (i) the cognitive system's information-processing architecture; and (ii) the state of the external world. It will take me a moment or two to arrive back at a more detailed discussion of these two factors. I need to begin with a few remarks on the very idea of a target.

According to the formulation adopted in this work, misrepresentation is a mismatch between what a representation bearer is *applied to* and what the representation bearer is *true of*. And, as Cummins writes: "The distinction...between what a representation is applied to and what it is true of is precisely the distinction between a representation's *target* and its *content*" (1996, 8; emphasis altered). Misrepresentation results when a representation bearer token fails to hit its target, when it is used to represent something to which its content does not correspond. While representation bearers stand-in for their representational objects, a use of a representation bearer "intends its target." That is, "the target of tokening a representation is, as it were, the thing the representation is intended to represent".[92] So, the idea of a target is "essentially a functional notion", in that it is the state of affairs the cognitive system "needs to represent" when it tokens a certain representation bearer (Cummins 1996, 6-8). A target is what a certain representation bearer token is *aimed at* (on a particular occasion of its tokening). As indicated in the earlier discussion, it is representation bearer *tokens*, not types, that have targets. A token of a certain representation bearer content-type can have different targets on different occasions, depending on the circumstances of its

[91] The reason why I have employed my own terminology in addition to that introduced by Cummins (1996) will become apparent subsequently. Up to that point, I'll mainly stick just to "targets", since I'll be primarily concerned with providing an exposition of Cummins' discussion.

[92] Not necessarily consciously intended, of course (Cummins 1996, 8).

tokening. Content does not determine target. As I've made plain, it is precisely the crucial extra degree of freedom afforded by having contents and targets determined independently that allows for the capacity for misrepresentation.

Let me turn to Cummins' (1996, 5-7) example of a chess-playing computer for a brief illustration. Chess computer *S* is involved in an endgame. It has a representation bearer which represents the current board position (P1). A subroutine LOOK-AHEAD examines the possibilities regarding the opponent's next move, and proposes a certain move in response. Let this potential sequence of moves be *M*. What LOOK-AHEAD needs to represent is the position after *M*, namely, position 2 (P2). It needs to token a representation bearer with the content "P2". Call such a representation bearer R_{P2}. Now, suppose that *S* tokens not R_{P2} but R_{P3}, which represents position 3 (P3) and not, as required, P2. *S*'s tokening of R_{P3} here is a misrepresentation. It is a misrepresentation because "there is a mismatch between the state of affairs *S needs to represent* when it tokens R_{P3}, namely P2, and the state of affairs RP3 *actually represents*, namely P3" (Cummins 1996, 6; emphasis added). The former state of affairs is *S*'s *target* in tokening R_{P3} on this occasion. The latter, of course, is its content. The tokening of a representation bearer is an instance of misrepresentation when the bearer's content does not satisfy the system's target on the occasion at issue.

P2 is *S*'s target in tokening R_{P3} because *S* is after a representation of P2 – a representation bearer with "the position P2, the position after *M*" [93] – to calculate its next move. P2 is what *S* is aiming to represent when it produces a representation bearer representing P3 (R_{P3}). And P2 is the state of affairs *S uses* R_{P3} to represent, the state of affairs to which R_{P3} is *applied.* As Cummins (1996, 7) writes: "Representing the position after *M* – P2 in this case – is, in short, the function of tokening R_{P3} on the occasion in question, even though, of course, R_{P3} does not represent P2 but P3." In order to understand why *S* (erroneously) tokens a representation bearer with content "P3" (when what it requires is a representation of P2), we would have to examine the processing episodes leading to *S*'s generation of R_{P3} at this time. Targets, of course, don't guarantee the tokening of representation bearers which hit them. Nor, as indicated, can we read the target off the content of the tokened representation bearer. In general, then, the determination of targets is a

[93] There is a subtle ambiguity in this expression used by Cummins; likewise, virtually the same expression in the next quotation. I'll return to this point shortly.

matter of "...the *representational function* of tokening a representation [bearer] on a particular occasion in a particular context...." (Cummins 1996, 7; emphasis added). In the chess example, the target of a representation bearer tokening at the time we're considering is P2 irrespective of the content of the representation bearer actually tokened.

On Cummins' account, the representational function of tokening a representation bearer reduces, for one thing, to the function of the *"intender"* which produces a representation bearer token at a particular time in a certain circumstance. An intender, that is, is a mechanism possessed by a cognitive system whose function it is to represent certain states of affairs. Cummins writes:

> Think of cognitive systems as incorporating mechanisms whose function is to represent certain things. For example, the function of a simple visual system might be to represent the local spatial layout – the relative sizes, shapes, and distances of objects from one another and from the observer. When this mechanism constructs a representation, the target of the representation it constructs is the current local spatial layout, whatever that happens to be (1996, 8-9).

Intenders, then, are cognitive mechanisms which have specific representational functions. Large, coarsely circumscribed intender mechanisms can have a number of embedded intender systems, that is, mechanisms whose representational functions are more narrowly circumscribed. While the visual system as a whole has the representational function of representing the local spatial layout, it contains more elemental intender mechanisms whose target domains include, for example, the representation of edges, colours, motions, and depths. The targets of representation bearer tokens are determined (in part) by the representational functions of the intender mechanisms that generate the representation bearer tokenings. *Mechanisms* for generating representation bearers (as opposed to the representation bearers themselves) are just the sorts of things that can have a *selectional history*, and are therefore ripe for the ascription of representational functions (recall the above criticism of Dretske's more recent TCD, Chp. Two, subsec. 2.3.).

So, one element in the determination of targets is the representational function of the intender mechanism which produces certain representation bearer tokenings. In the chess example, the chess computer contains a mechanism whose function in tokening R_{P3} is the position after *M*. That is, the *mechanism's* target in tokening a representation bearer is the position P2, the position after *M*, whatever that position in fact happens to be. Or consider a (crudely specified) mechanism of the visual system whose representational function is to represent the current object of foveation ("the object currently before me"). The target of this mechanism is the current object of foveation, whatever the object actually *is* (a horse, a cow, a bovine shaped bush, a Holden Monaro, or whatever). The component of target fixation which is concerned with intender mechanisms plainly yields only a coarse-grained conception of a target, one not specific enough to legitimately talk of a comparison between content and target. However, as indicated, there is another element to the determination of targets, namely, the current state of the external world (see Cummins 1996, 8-9, 118-119). Target determination here is *indexical* in character. A representation bearer token generated by an intender mechanism is aimed at a certain variable; the world's contribution to target fixation is, so to speak, to attach a value to the variable. So while our chess computer has an intender whose target is the current position (whatever the current position may be), it's the state of the external world that determines what the current position actually is. Or again consider an intender whose representational function is to represent the current object of foveation. The variable which is the mechanism's target is the current object of foveation (whatever value that variable exhibits). The state of the world determines the value of that variable; it fixes a determinate target. It is only with the conjunction of these two elements of target determination that we can make an evaluation of the veridicality of a representation bearer tokening. Say an intender mechanism with the target of the current object of foveation generates a certain representation bearer token, one with the content "horse". This representation bearer token is applied to the current object of foveation. But in fact the actual object residing in the world at the point of foveation is not a horse but a cow. The target – jointly fixed by the function of the intender mechanism and the state of the world – is, then, a cow. The cognitive system applies a representation bearer token with content "horse" to the target "cow". Content fails to coincide with application, with target, and so we have an instance of misrepresentation.

I can now reveal why I have engaged in the practice of employing the term "intentional object" in addition to "target" as coined by Cummins. So far I've used the two terms synonymously, but I now want to explain that they are only synonymous on a particular rendering of "target". I outlined above that there are two components of target determination, the representational function of an intender mechanism, and the state of the external world. There is, however, a subtle ambiguity in Cummins' use of "target" in his formulations of misrepresentation. Cummins uses "target" to refer to *both* the function of an intender mechanism and the particular state of the world at which a representation bearer tokened by the intender is aimed. But, as shown above, we cannot usefully evaluate the veridicality of representation when target is understood in the former sense. Again, the target of an intender is a certain variable (for example, the position after *M*). The function of such a mechanism is silent on the value of that variable (its function is to generate representation bearers which represent that variable, whatever the position after *M*, for example, actually is). As yet, it makes no sense to ask after the veridicality of representation. The question of whether a token representation bearer is veridical or in error is only legitimately framed when it involves a comparison between the value of a variable which constitutes the bearer's content and the value of a variable which constitutes the bearer's target. Part of the problem here is that Cummins vacillates between talk of the *system's* target and the *representation bearer's* target. But these two are not equivalent, and, as indicated, it must be very clear which notion of target is being invoked in the formulations of misrepresentation. The distinction here is legitimate; the two notions of target correspond, of course, to the two components of target fixation. We can, I think, legitimately talk of the representational function, and so target, of the system or *mechanism* (for example, the position after *M*, or the current object of foveation), and the representational function (on a particular occasion), and so target, of a representation bearer *token* (for example, what P2 actually is, or cow). However, as suggested, it's the latter which must be in place in order to justifiably talk of an instance of misrepresentation. Only the two components of target fixation *together* yield the phenomenon of misrepresentation.

So, I propose to stipulate and adopt the following terminological convention. I call that notion of a target which is constituted by the representational function of an intender *plus* the world's contribution to target fixation the *"intentional object"*. The intentional object is a *value* of a variable (as opposed to just a certain variable, whatever its value may actually be). And misrepresentation is a mismatch between the values of a variable (or values of variables): the value of a variable which is a token bearer's content, and the value of a variable which is the token bearer's intentional object. On my formulation, then, misrepresentation is the failure of representational content to coincide with the intentional object.[94]

Intentional object determination, on the basis of the above discussion, comes down to the following: *IO* is the intentional object of an application of a representation bearer token insofar as it is the function of that tokening of the bearer to represent *IO*.[95] Though the issue of function attribution is a large and contentious one that extends beyond the philosophy of mental representation, it is worthwhile examining what Cummins has to say on the analysis of functions in regard to intentional object determination.

There are two general kinds of functional analysis: selectionist theories and design theories. Selectionist theories of functions reduce functions to adaptiveness and, ultimately, selectional history. It is the function of x to do f because doing f conferred a survival advantage, and so items of type x were replicated because (on at least some occasions) they performed f. Design theories unpack the function of an item by reference to its functional role, "that is, in terms of its contribution to some capacity of a containing system." So, "the (or a) function of x in a system S is f relative to a capacity C of S just in case S's capacity C analyzes (in part) into x's capacity to f" (Cummins 1996, 114-116; notation altered).

Now, as Cummins dutifully recognises, there is a significant problem that must be overcome if the standard theories of functions are to underwrite intentional object fixation.[96] Both selectionist and design theories attribute functions to *types* of states, objects, or mechanisms. For example, Millikan's teleosemantics seeks to explicate content determination in terms of a type of representational item's Normal function, that which contributed to the replication of the type of item. But as I've made plain, it is a tenet of the alternative approach that attributions of representational function pertain to representation bearer tokens, not types. Tokens of the same representation bearer type can be applied to different intentional objects on different occasions. The idea that content determination concerns representation bearer types while tokens of the same content-type can have disparate intentional objects is at the heart of the alternative approach put forward above. A selectionist

[94] So, of course, Cummins' formulations of misrepresentation are not incorrect. However, they are at their most precise when "target" is portrayed (as is often the case) in the sense of the representational function of an intender mechanism plus the state of the world.

[95] This formulation is adapted from that put forward by Cummins (1996, 113).

[96] Actually, Cummins (1996, Chp. 8) detects two problems, but I'll confine attention to that difficulty which is the most significant and readily apparent.

theory, then, does not look a likely candidate to underwrite intentional object fixation because a particular representation bearer token does not have a selectional history, and understanding the function of representation bearers in terms of the function of the type to which they belong blocks the route to the account of misrepresentation defended in this work. As Cummins (1996, 116) notes, design theories likewise clearly attribute functions to types of items (states, objects, or mechanisms). But to worry about both selectionist and design theories on the grounds that they attribute functions to types of items is to say that the required theory of functions is one which attaches representational functions to representation bearer *tokens*. This looks problematic because it obviously entails that for every intentional object at which a cognitive system can aim its representation bearers, there must be a mechanism (an intender) for generating representation bearer tokens whose function it is to represent each particular intentional object. And since complex cognitive systems like humans may possess an unbounded number of intentional objects, we are forced to posit an unbounded number of intenders, which appears totally implausible.

However, the above difficulty is not insurmountable. There are two considerations which can be wedded to the standard analyses of functions in such a manner as to secure the required unbounded number of intentional objects (Cummins 1996, 118-120). The first has been encountered already. It was proposed above that there are two components to the determination of intentional objects: (i) the representational function of the intender mechanism that generates representation bearer tokens; and (ii) the state of the external world. Now, the crucial factor here is that intentional object fixation typically has an *indexical* nature. In sensing their environments, cognitive systems generate representation bearers that stand-in for *current* stimuli. That is, what was left implicit in the earlier discussion is that the representational function of an intender mechanism ranges over not just a certain, rather specific variable, but also the state of that variable at a *particular time*. So, the chess computer has intenders for the *current* board position, the most likely *next* opposition move, the position *after M*, and so on. And a (coarsely individuated) visual system has, for example, an intender whose representational function is the *current* object of foveation. It is then the actual state of the world that serves to complete the fixation of determinate intentional objects. That the current board position is P1 (white king at 3,7, white bishop at 2,6, and so on) makes P1 the intentional object of a representation bearer token generated by the chess system's "current position" intender. That it is a cow located in the external world at the cognitive agent's point of current foveation makes a cow the intentional object of a representation bearer tokened by the agent's "current object of foveation" indender. The sensory systems of behaviourally versatile cognitve creatures must possess a host of such plastic indexical intenders. A surfeit of examples could be generated; a few include: the current local spatial layout, the timbre of the voice presently being heard, the phonetic (syntactic) structure of the current verbal input, the far edge of the surface now being traversed, the current hypothesis, the name of the person just introduced, the speed of the creature now approaching, the current pressure on my arm, the last word in the sequence just completed, the tannin level of the shiraz just swallowed, the colour of the fruit about to be picked, the sudden degree of variation from the vertical of my body, and so on. The crucial point is that the indexicality of intenders means that while a cognitive system has a finite number of intenders, it can nonetheless have an unbounded number of intentional objects. Any state of affairs the world is capable of exhibiting is a potential intentional object for a cognitive system possessing an intender mechanism with the function of representing variations in that state of affairs.

The second factor which, according to Cummins, permits an unbounded set of intentional objects concerns the notion of embedded, or as he says "nested", intenders. The idea is that intenders come in various resolutions, with those having more fine-grained or specific representational functions embedded within intenders possessing more coarse-grained or broader representational roles. Such embedded intenders interact with one another, with the tokening of representation bearers in certain intenders bringing about tokenings in others. Consider again Cummins' chess system example. The chess system has an intender whose representational function is the position after *M* (*M*, recall, is a certain finite sequence of moves). In order to operate, such an intender must have representation bearers standing-in for a move sequence. These representation bearers are themselves generated by another intender mechanism, one whose job it is to generate representation bearers standing-in for move sequences. The latter intender is embedded within, and interacts with, the former. The intender that tokens representation bearers standing-in for move sequences, in turn, presumably produces them from representations of moves. Supposing that a move is represented in the form of an ordered pair of board positions, move representations are, at an even deeper level of embedding, produced from those representation bearers which are proxies for board positions (Cummins 1996, 119).

So, the idea of nested intended mechanisms, in concert with the claim that intentional object fixation is typically indexical in character, shows that the requirement for an unbounded number of intentional objects can be plausibly met. The alternative proposal for the accommodation of misrepresentation may have the structural isomorphism TCD at its heart, but a good deal of its machinery, in contrast to the orthodox approach, lies beyond the account of content determination itself. Finally, note how the world's contribution to application, to intentional object fixation, reveals why the orthodox approach has *deep* difficulties with the phenomenon of misrepresentation. Even before I introduced some of Cummins' ideas on intentional object determination, there had already been some indication of the role of the external tokening object in application. At a number of points I've remarked that, in the perceptual case, it is typically the external tokening object to which a representation bearer is applied. Any account of misrepresentation cannot deny that the external world has some part to play in application. And, as I've previously highlighted, the orthodox TsCD in some way or another accord the crucial content determining role to the world, the external tokening object. Content determination and application, on the orthodox approach, inextricably overlap. A viable means for the accommodation of misrepresentation, however, as I've been at pains to demonstrate, must have content and intentional objects separately determined. And structural isomorphism, as far as I can see, is the only TCD which can meet the demands of such an alternative proposal for misrepresentation. If the line of argument pursued to this point is near to being on track, structural isomorphism is not just another candidate of equal status to the prevailing answers to the question of content determination, it is to be preferred over the prevailing TsCD.[97]

---

[97] I should again emphasise that by "prevailing" TsCD I mean already existing TsCD, which includes functional role semantics (FRS). I have not explicitly argued that FRS cannot explain misrepresentation. However, as Cummins (1996) has shown, FRS is also a "use" theory. It thereby falls within what I've called the "orthodox approach" to content determination and misrepresentation, and so (as Cummins has also argued) cannot solve the problem of misrepresentation. My conclusion here, then, is that only the structural isomorphism TCD can satisfactorily handle misrepresentation.

CHAPTER FIVE

INTERNALISM, EXTERNALISM, AND THE NOTIONAL WORLD

The overall aim of this chapter is to argue for an internalist theory of content determination of the kind afforded by structural isomorphism, and, concomitantly, against the standard externalist approach. Structural isomorphism, to reiterate, is an internalist account of content *determination* because the content conferring objective relation is internally instituted, that is, put in place by the brain. According to the theory, it is with the organisation of neurophysiological states into a physically structured system – one which aligns itself with a system of values of an external variable – that the neural system accrues the capacity to represent the external variable. The case for internalism presented herein centres on certain more or less empirically based results from two domains of human representational activity, namely, temperature and colour (secs. 1. and 2., respectively). The investigation into temperature and colour will, on the one hand, emphasise the *structured* nature of our representational experience, and, relatedly, reveal the difficulties in seeking to explain the character and content of representational experience in the prevailing externalist manner. The main line of argument, that is, divides into two more particular, overlapping concerns.

One task will be to illustrate the plausibility of the kind of representing systems proposed by the structural isomorphism TCD. I will employ the domains of temperature and colour sensation as examples of the structured nature of our representational experience, and thereby argue that such examples are strongly suggestive of structural isomorphism's claim that our representational capabilities are subserved by proprietary, physically structured systems of representation bearers. The structured nature of our representational experience, like the selectivity of mental representation discussed above (Chps. Three and Four), is a characteristic of our mental representational capabilities that, to my thinking, has received insufficient attention in recent philosophy of mind.[98] In the more prominent discussions of the orthodox TsCD, we find it virtually ignored. Dretske (1995, 21), as a rare exception, points out that there is a certain kind of structure "in our

[98] See, however, Austin Clark (1993) for a nice example of taking such structure seriously.

experience of colour and sound", but does little more than touch upon the idea. Moreover, it goes unremarked that such a claim is three-way ambiguous, between: (i) a represent*ing* system under a *physical* (computational, or neurophysiological) interpretation; (ii) a represent*ing* system under a *phenomenological* description; and (iii) a represent*ed* system *qua* system of represented values or *contents*. One aim of the investigation into temperature and colour will be to link together (i) and (ii) in the shape of dedicated, physically structured representing systems. That is, the representing systems of the structural isomorphism TCD sit comfortably with, indeed *explain*, the relationships holding among a system of representational experiences.

A further concern will be to argue that the empirically flavoured results strongly mitigate against the orthodox conception of content and the standard externalist approach to content determination. Such results, however, are very suggestive of the internalist structural isomorphism TCD. That is the more readily comprehensible of the two related tasks which fall within my second concern. The more demanding task I undertake is showing how the structural isomorphism TCD forces a significant reconception of the represent*ed* world. The world of mental content, I will argue, is a *notional* world, rather than the world itself (the objective world). Mental content is not the representation of *straightforward* features of the world itself, rather it is an averaged, contextualised, "slice" across the world, in some sense a cleaned-up "construction" of the brain's representing systems. Finally, and in accordance with both internalism and the idea of a notional world, I put forward what I call a *"projectivist"* theory of intentionality (sec 3.).

The overall hope, then, is that (i), (ii), and (iii) above be tied to one another under the rubric of the structural isomorphism TCD. Relatedly, the present chapter further develops and defends the "directedness" conception of intentionality first outlined in Chapter Three.

1. The Representation of Temperature

One capacity a TCD pitched at the representational base should explain is the representation of surface skin temperature. Temperature sensations point beyond themselves to temperature states of affairs at various locations on the skin surface. As Akins (1996, 3-12) suggests, one possible account would suggest that the thermoreceptive system is thermometer-like. On this view, the skin's temperature receptor system is calibrated to respond to *objective* skin temperature, such that there is a constant function from objective temperature to the states of a single, continuously variable thermomechanism. Each particular state of the receptor system covaries with – is tokened by – a particular temperature state. That is, the receptor system responds with a unique signal which correlates with a particular, objective temperature.

Such a proposal for the representation of temperature exhibits many characteristics of the orthodox approach to content and content determination (Chps. Two and Three above). Neural states tokened by temperature stimuli are indicators, or detectors, of objective temperature. Mental representation bearers stand-in for an objective feature of the world, in this case, objective temperature. Objective temperature is, of course, the actual temperature instantiated in the world (at the skin surface), in contrast to the "felt" temperature as represented by temperature sensations. Temperature content is the representational object itself. I will refer to this conception of temperature content as "external temperature" or "objective temperature". Again, the strategy is to begin with veridical content, that is, correct application. A type of representation bearer has as its content that particular objective temperature which, in the special content-determining circumstance, occasions tokens of that particular bearer-type. And, of course, content determination is seen in terms of an externally-instituted relation between the world itself (objective temperature) and the neural bearers of temperature sensations.

What I have portrayed as the orthodox or "connectedness" conception of content and content determination has some affinity with a conception of sensory systems Akins (1996) terms the "traditional view". Akins' argument (of which the temperature example is a component) is that this view is false. And while I am sympathetic to her overall line of argument, my focus will be with those findings Akins (drawing on work by Hensel, 1982) employs to demonstrate that temperature perception does not conform to the proposal sketched above. Indeed, I will go further than Akins and suggest that the temperature case provides support for the internalist account of content determination put forward by structural isomorphism.

The proposal outlined above suggests that temperature perception is subserved by states of a single thermomechanism operating in a uniform manner to indicate objective temperature. However, right from the earliest stages of processing – the thermoreceptive system of the skin – temperature perception does not conform to this picture. Rather, temperature sensations are initiated by the activity of four different types of receptors: two thermoreceptors, "warm spots" and "cold spots", and two pain receptors, "nociceptors" (active only at the very upper and lower temperatures). Between the upper and lower extremes, temperature stimuli produce activity in either cold receptors or warm receptors, though there is a small range of temperatures (at the upper end of the temperatures to which the cold receptors respond) where activity is produced in both types of thermoreceptors. Moreover, temperature sensations are a function of the particular ratio of cold to warm spots at various locations on the body. Overall, there are many more cold receptors than warm receptors, though the ratio varies from one part of the body to another. (For example, there is a ratio of 8:1 cold to warm spots at the nose, 4:1 at the chin and cheeks). The differing ratio of cold to warm spots at various bodily locations means that the *same external* temperature can give rise to *different* temperature *sensations*. Even at this early stage, then, trouble is brewing for the orthodox view.

The two types of thermoreceptors respond to both constant (external) temperatures and temperature changes. Each receptor's activity under stable temperature conditions is its *static function*, while its response to temperature change is its *dynamic function* (see Figure 5.1.). The static function is displayed by a curve showing the neuron's firing rate at each (external) temperature within its response range. The response of both warm receptors and cold receptors is nonlinear, though the two receptor types have different response profiles. The warm receptor begins to fire at a certain temperature, and responds to higher temperatures with increasingly higher firing rates until its maximum response temperature is reached, above which firing immediately ceases. The cold receptor responds to a wider range of temperatures, and it gradually fires more frequently as temperature increases until the midpoint of its response range, gradually decreasing again until the highest temperature at which it responds is reached.

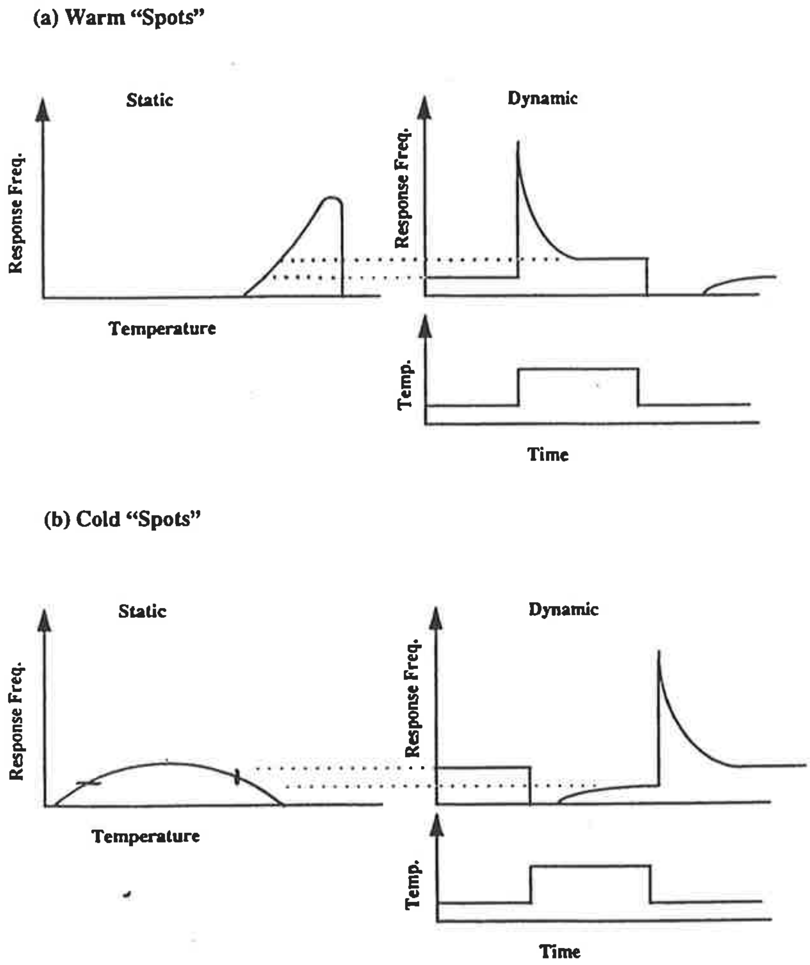

Figure 5.1. The static and dynamic functions of the warm and cold receptors. To depict dynamic function, both receptors were exposed to a sudden temperature increase followed by a sudden temperature decrease. The warm receptor exhibits a dynamic response to temperature increase alone; the cold receptor exhibits a dynamic response to temperature decrease alone. (From Akins 1996, 347).

Here I highlight the first of three important points: "*The static functions of neither the warm spots nor the cold spots are thermometer-like, with a certain set increase in firing rate per degree of [external] temperature change*" (Akins 1996, 347; my emphasis). In other words, there is not a constant function from external temperature – the *objective* temperature of the skin – to activity levels of the skin's receptor system (the transducer system). Systematic proportional differences in objective temperature are not reflected in proportionally equivalent differences in the activity levels of the thermoreceptors. So differential activity levels of the thermoreceptors at different times are not to be explained by some constant correlation with objective temperature differences. This finding, along with the fact that temperature sensations are a function of the ratio of cold and warm receptors at various bodily locations, undermines the orthodox "indicator" (or thermometer-like) view of temperature content determination.

The second important point is that *any given thermal sensation does not necessarily correlate with a certain external temperature* (Akins 1996, 351). That is, there is also no constant function from external temperature to temperature *sensations*. The same external temperature can token different temperature sensations; or, the same type of temperature sensation can be tokened in response to different external temperatures. The reason is that temperature sensations do not arise directly from the functional properties of individual thermoreceptors. Rather, temperature sensations are the product of a collective neural response. The overall pattern of activity across a neural population is partly a function of the absolute number and ratio of the two receptor types (see Akins 1996, 351-352). Again, such a result clearly looks to mitigate against the orthodox view that particular instantiations of objective temperature determine content by virtue of occasioning particular tokenings of representation bearers. The orthodox view is further undermined by the behaviour of the thermoreceptors in response to temperature *change*. Akins writes:

> The felt change in temperature for a specific temperature change will depend upon the starting temperature of the skin. If the temperature of a warm spot is increased at the bottom of its response range, the dynamic burst will be small; if it is warmed at the top of its response range, the burst will be very large (1996, 351).

The dynamic function of both warm spots and cold spots is *context dependent*.

So, neither the behaviour of the thermoreceptive system, nor temperature sensation, correlates with objective temperature. In both regards, we are concerned with the relationship between temperature representation bearers and objective temperature. The physical properties of the thermoreceptive system (such as the firing frequencies of its constituent neurons), and the qualitative properties of temperature sensations, are, of course, two different kinds of description of temperature representation bearers. Such representation bearers stand-in for (have as their contents) temperature states of affairs in the "external" world (surface skin temperature). Differences in those representation bearers (whether phenomenological or physical/neurophysiological/computational) correspond to differences in represented temperature, temperature content. But since the properties of temperature representation bearers – under either a physical or phenomenological interpretation – do not correlate with objective temperature, temperature content determination cannot be a matter of reliable correlation between a certain objective temperature and the tokening of a representation bearer of particular type.

So we are left with two questions. First, if external temperature does not correlate with temperature sensation, what determines felt temperature? Second, if external temperature does not – via some kind of indication relation – determine temperature content, then what does determine the content of temperature representation bearers? I've already touched on the answer to the first question. Temperature sensation is a product of the constitution and behaviour of a neural population at a certain bodily location. Felt temperature, that is, is determined by the overall pattern of activity across a coalition of warm and cold receptors. While external temperature (standardly) elicits activity in a neural population, the resultant overall pattern of activity is a product of the total number and ratio of the two receptor types (and, with temperature change, the starting activity profile). [99]

---

[99] Such an answer may look to be trivially true. After all, given materialism, temperature sensation and neural activity are not independent *relata*, but simply two varying descriptions of the same, ultimately physical, state of affairs. Temperature sensations are instantiated by neural events, so of course the character of any given temperature sensation coincides with the character of neural substrate. But there is a significant difference between claiming that the particular character of such an internal state of affairs is straightforwardly determined by external or objective temperature, and proposing that the constitution of a neural population "goes to work on" the objective temperature stimuli (such that the nature of the resulting temperature sensation is influenced by the make-up of the neural population).

The story of temperature sensation obviously doesn't end with the behaviour of the thermoreceptive system of the skin. Signals from coalitions of receptors at the skin surface are sent to the brain, where temperature sensations are instantiated. Much less is known about the nature of the processes here than the operating principles of the transducer thermoreceptive system. However, it is of course a central conjecture of the structural isomorphism TCD that temperature perception is subserved by a temperature-specific representing system, one which admits of a systematically structured organisation along certain physical relations holding among neural representation bearers. And while, as demonstrated, the value-structure of external temperature may not correspond to the value-structure of temperature sensations, it's the claim of structural isomorphism that such a physically structured representing system *qua* neural system nicely comports with the *structured* nature of temperature sensation. The range of temperature sensations we are capable of experiencing exhibits a salient phenomenological structure. Within those boundary points which implicate the activation of the pain receptors, the continuum of temperature sensation – from uncomfortably hot through pleasantly warm, tepid, unpleasantly cool, to bitterly cold – manifests a similarity structure, a set of similarity relations. Certain temperature sensations are more closely phenomenologically related than others (a pleasantly warm temperature sensation is more like a mildly uncomfortably hot sensation than a bitterly cold one). According to the structural isomorphism TCD, such phenomenological structure is no accident: the range of temperature experiences coincides with a range of dedicated neural items which also exhibits a set of similarity relations (along a continuously variable property of the neural substrate). The physically structured, domain-specific system of neural representation bearers *accounts for* the phenomenological structure exhibited by the system of temperature sensations. So, the claim here is that the structure of temperature sensation is not determined by the range of external temperature stimuli, rather it is the product of the kind of physically structured representing system posited by structural isomorphism.

I turn now to the second question, the determination of temperature content. I've already indicated that the empirical findings suggest temperature content determination does not conform to the prevailing externalist explanation. However, what I isolate as the third main finding in Akins' discussion initially looks to also be in tension with the structural isomorphism TCD. Akins writes:

> ...thermal sensations as a whole do not reflect the structure of thermal stimuli, whether stimulus T1 is greater than, less than, or equal to stimulus T2. The water, as we wade into it, initially lessens the skin temperature of each body part about equally; but it certainly does not feel that way. Some parts feel much colder (1996, 351).

The fact that the same external temperature can produce different temperature sensations at different bodily locations is of course due to the above explanation of the second finding: the character of any given temperature sensation is determined by the collective activity of a neural population, and such an overall pattern of activity is itself influenced by the number and ratio of cold and warm receptors at particular bodily locations. I also take it that Akins does not just mean that the same external temperature can token different temperature sensations (at differing bodily locations), or that different objective temperatures can, at differing times, token the same kind of temperature sensation. That would be to assert the independence of content determination from the tokening relation, something which, of course, I've argued for at length above (though Akins does not articulate such a view, nor formulate a use-independent TCD). Rather, Akins' point is that the phenomenological structure of a system of temperature representation bearers does not mirror the structure of variations in *objective* temperature. Variations in objective temperature do not line-up with variations in the "values" of temperature sensations. And since, as argued above, systematic variations in temperature sensation (the representing system on a phenomenological interpretation) are explained by corresponding systematic variations in their neural instantiation (the representing system on a neural interpretation), there obviously looks to be trouble here for a theory that proposes that content determination is a matter of a structural complementarity between an external variable and a system of mental representation bearers.

It is here that I must signal an important shift in the character of the TCD being developed in this work. Or, more precisely, I should signpost the integration of certain views developed earlier in the piece (particularly Chp. Three) with the current continued formulation of the structural isomorphism TCD. Specifically, and in line with "representational-object-*as-represented*" conception of mental content proposed above, the representing systems of the structural isomorphism TCD do not straightforwardly reflect the structure of the world *itself*, in this case objective or absolute temperature. Rather, the represented world is a *purported* or *notional* world, "a way" the world itself might be. The represented world is a way that mental representation bearers represent it *as* being. According to the structural isomorphism TCD being developed here, there may not be an isomorphism between a system of temperature representation bearers and objective temperature, but there *is* an isomorphism between a dedicated representing system and temperature *as-represented.*

So, while representation bearers with internally determined content certainly stand-in for values of an external variable (values the variable may be able to manifest), they do so not by mirroring the structure of the variable *itself* (the objective structure), but rather by *"imposing"* a structure upon the external variable. The internal structure of a representing system takes a structured "slice" of an external variable, such that the former's value-structure is not isomorphic to the latter's objective value-structure, but the value-structure as represented by a representing system. The physically structured representing system is structurally isomorphic to a purported or notional world, not the world itself.[100]

The notional world, however, is not completely divorced from the world itself. There will typically be a certain, even significant, "overlap" between the objective and notional worlds. Moreover, different representing systems may overlap with the objective structures of their representational domains to varying degrees, and the one representing system may overlap its objective represented variable to differing extents at different times. The situation with temperature provides some illustration of these points. While the value-structure of objective temperature does not map straightforwardly onto the value-structure of temperature sensations, it is in general the case that higher objective temperatures roughly line-up with higher felt temperatures. In addition, as Akins (1996, 351) reports, the static function of the warm receptor does display some kind of correspondence with a certain range of temperatures, in that "lower temperatures do elicit lower firing rates...." Nonetheless, whatever the exact relationship between a notional represented feature and the feature itself in the objective world, I maintain that mental content coincides with the former. The feature as-represented is a *"skewed"* representation of an external variable. The reason, as suggested, is that a

[100] The discussion of this section is only a first pass at the notional world. The idea is discussed again in the following two sections.

representing system in some sense imposes a structure on a represented feature. It "carves up" the represented variable according to its own physical structure. That is, the structure laid upon a represented variable – that which generates the variable as-represented – is determined by any "skewing" of the structure of the representing system. Consider the following passage from Akins:

> ...the thermoreceptive system *embroiders* its account of the temperature states of the world. ...[I]t is probably unfair to say that this particular system actually manufactures fictions..., but it does appear prone to chronic exaggeration. At the lower and upper limits of response for the cold spots and warm spots, respectively, a small temperature change elicits a hysterical response. Moreover, given the differing ratio of cold spots to warm spots, no body part *simply reports* its surface temperature. Each body part exaggerates its own state in accordance with its own interests and sensitivities (1996, 352; my emphasis).

The thermoreceptive system does not slavishly record objective temperature. This system, and the operation of sensory systems in general, is, as Akins (1996) nicely puts it, *"narcissistic"*. A representing system is "skewed" in a manner which reflects the organism's constitution and environment, and to serve the creature's needs and interests. The structure of the temperature representing system is tuned to meet narcissistic concerns in that the most dramatic skewing occurs at those temperatures ranges which are potentially *dangerous* to the organism. It is an interesting exercise to speculate on what skewing may be present in the representing systems of other creatures, say, polar bears.

In my view, then, the empirical results Akins marshals are highly significant. They provide telling evidence against the prevailing externalist strategy of explicating content determination. And they are strongly suggestive of the "representational-object-as-represented" conception of content I proposed earlier in this work, a conception now filling-out in the shape of the represented world as a notional world. The significance of the negative strand of Akins' discussion notwithstanding, she is, however, without an alternative picture of the determination of content. Part of the problem here is that despite her insistence that it is the cumulative activity of a neural population that corresponds to temperature sensation, I think Akins does at times place too much emphasis on the behaviour of individual warm and cold receptors. This may be appropriate when providing a critical examination of the orthodox approach, as indication occurs early in the processing chain (and to be fair to Akins, her main undertaking is to undermine the traditional, indicator view). But the structural isomorphism TCD shifts the focus from *transducer* systems to the *brain's representing* systems. And it proposes that the representing domain involved in the structural isomorphism relation consists of a system of activation patterns generated across neural networks.[101] Such a set of patterns of activity exhibits a similarity structure in terms of the patterns' profiles or "shapes". The overall shape of a pattern of activity is determined by the nature of the signal sent from the transducer system, itself a function of the number, ratio, and relative activity levels of both cold and warm receptors making up the transducer system. The structural isomorphism TCD is, of course, a *hypothesis*. But as far as I can see, there is nothing in Akins' discussion which shows that the structural isomorphism TCD – with the isomorphism holding between the domains I've lately rehearsed – is an untenable hypothesis. Indeed, I think the findings she reports are highly suggestive of such a hypothesis. Interestingly, in her characterisation of the orthodox view, Akins touches upon the kind of representing system posited by the structural isomorphism TCD. Criticising the prevailing indicator account, she writes: "...constant correlation is not enough. ...the representational relations among the sensory signals must mirror the relevant relations in the sensed domain" (1996, 343). I think that Akins is right to insist that standard view of mental representation bearers as indicators is inadequate. But at that point in her discussion where she considers the kind of representing system proposed by the structural isomorphism TCD, Akins portrays structural isomorphism as necessarily in keeping with the orthodox view. This need not be the case. The structural isomorphism TCD I favour has a representing system structurally isomorphic to an external variable *as-represented*. And this shows just how significantly the account of content determination emerging here deviates from the orthodox understanding.

2. The Representation of Colour

This section deploys colour vision to the same ends as the above discussion of temperature perception. The discussion here is significantly more detailed than that

[101] The terminology here, of course, suggests a connectionist perspective on the brain's representing systems. I wed structural isomorphism to connectionism in Chapter Six below.

of temperature, so let me first outline the argument. I begin by showing that an explanation of colour is not to be found, as philosophers have often supposed, in the external world itself, in, for example, the physical characteristics of light (subsec. 2.1.). Accordingly, I turn attention to the organism, specifically, the operation of the transducer system at the retina and, second, those processes dealing with outputs from the receptor system. Of particular importance here is the contemporary "opponent process" theory, now largely acknowledged as lying at the heart of current scientific understanding of chromatic vision (subsec. 2.2.). I next examine the structure of colour phenomenology, the set of relationships obtaining among a system of colour experiences depicted by a colour *quality space*. I spend some time detailing the nature of this structure, and go on to show how it is explained by opponent process theory in a manner concordant with the physically structured representing systems of the structural isomorphism TCD (subsec. 2.3.). The internalist explanation of colour phenomenology and the contents of colour experiences is further pursued with particular focus on a *neurophysiological interpretation* of colour quality space (subsec. 2.4.). I then describe two considerations which, to my mind, show that the world of colour corresponds to a *notional world*, rather than the world itself (subsec. 2.5.).

Given the detailed nature of the discussion of this section, a brief overview of chromatic vision – from the transducer system of the retina to the character of colour experience – is also appropriate. The human visual system is sensitive to electromagnetic radiation of wavelengths between approximately 400 and 700 nanometres (nm). The receptor system of the retina consists of two subsystems, rods and cones, specialised for different tasks. The rods are specialised for dim (night-time) light levels, and the cones for daylight. Activity in the rods produces a rather impoverished kind of vision, which is also purely achromatic (only black, white, and grey perceptions result). Activity in the cones results in both achromatic and chromatic perceptions. The outputs of the cones are processed in a manner described by the opponent process theory of colour vision. The outputs of opponent processes comprise those neural signals sent from the retina to the visual cortex, with activity therein ultimately realising the rich colour experience the majority of us enjoy.

Colour experience consists of three distinguishable dimensions: hue, brightness, and saturation. These three dimensions are the differing respects in terms of which colour experiences are phenomenally related. If two colour experiences have the same hue, same saturation, and same brightness, then they are the same type of colour experience; they have identical qualitative character. *Hue* is the dimension of colour to which people refer in the everyday use of colour terms. (People will say "bright green" or "pale blue" to express the other dimensions). It is what makes following snooker much easier with a colour, rather than black and white, television. So, the hue of a colour is its redness, or green-ness, or yellowness, or blueness. White, black, and the greys – the achromatic colours – have zero hue. The *saturation* of any hue is the strength, or purity, of that hue. Colours with the same hue may be weakly hued (and thus close to grey), strongly hued, or somewhere in between. Variations in *brightness* range from black through the greys to white (see, for example, Hardin 1988, 25-26). The three distinguishable dimensions of colour, and the range of values within each dimension, mean that we can think of colour experience in terms of a *quality space*.[102]

2.1. Colour and the Physical Characteristics of Light

Some philosophers (most notably perhaps Armstrong 1978) have been tempted to claim that distinct colours are identifiable with distinct wavelengths of light. Such an identification, however, is unlikely to succeed. For one thing, perceived colour is a product of the intensity of light, the wavelength of light, and the relative proportion of broad-band radiation ("white light") mixed with a monochromatic light. So, when either the intensity, wavelength, or proportion of white light is changed, brightness, hue, and saturation all change, resulting in a different perceived colour (Cornsweet 1970, 234).

Perhaps, then, each of the three distinguishable dimensions of colour could be identified with some particular characteristic of light. *Prima facie* this looks more promising, since the dimension which changes the most with variations in intensity is brightness, while hue changes the most with variations in wavelength, and saturation changes the most with variations in the relative proportion of white light (Cornsweet 1970, 234). However, this proposal won't work either. First, as before, the three distinguishable dimensions of colour are not strictly independent.

---

[102] Good sources of information on chromatic vision include: Austen Clark (1993); Cornsweet (1970); Hardin (1984a, 1984b, 1985, 1988, 1990); Hubel (1988); and Hurvich (1981, 1985).

That is, a change in any one of the three physical characteristics of light produces a change in all three dimensions of perceived colour. Moreover, if we focus on pairs of perceived dimensions and the alleged corresponding physical characteristics, we find that the degree of covariation of one dimension with another is different for different pairs. Consider hue and saturation. Yellow, for example, is intrinsically less saturated than green. That is, a 580 nm light will be perceived as less saturated than a 503 nm light even though, in both cases, the relative proportion of white light mixed with the monochromatic light is the same (Hardin 1984a, 494). So hue and saturation can covary with differences in wavelength and relative proportion of white light held constant. Consider now the relationship between hue and brightness. Differences between hues are affected little by variations in brightness (intensity). That is, hue discriminations are easy and reliable under changes in brightness. However, brightness differences are significantly affected by variations in hue (wavelength). That is, brightness discriminations are difficult and less reliable when they have to be made with a range of different hues (Cornsweet 1970, 235).

Despite the complexities outlined above, some are still drawn to an attempt to identify distinct colours with distinct wavelengths of light. Perhaps this is partly because hue is the most salient dimension of colour experience, and hue is in some fashion related to wavelength. Offering some support in this regard is the existence of the "unique" hues: red, green, yellow, and blue. A unique hue is a hue which has no other chromatic component. So, unique yellow is the yellow hue with no red or green component, unique red is the red hue with no yellow or blue component, unique blue is the blue hue which is neither reddish nor greenish, and unique green is the green hue which is neither yellowish nor bluish. For most subjects under standard conditions, unique yellow is close to 580 nm in the spectrum, unique green near 500 nm, unique blue around 175 nm, while unique red can be produced by mixing a 700 nm red from the spectrum with a small amount of 475 nm light (Hardin 1985, 36). For any given subject, these results are very robust across different experimental sittings. There are relatively small differences among different individuals. Interestingly, the differences involve the whole structure of colour experience – the various colour experiences and their relationships – being uniformly shifted relative to other perceivers. (I will come back to the structure within the range of phenomenal hues subsequently). Hues other than the unique hues are all "binary" hues, perceptually consisting of a percentage of each of two unique hues.

The fact that unique red cannot be evoked by a particular wavelength of light should alert us to a difficulty in attempting to identify distinct colours with distinct wavelengths of light. Indeed, the situation is far more dramatic than the case of unique red suggests, and the identification of hue with wavelength is actually the most problematic of the attempted identifications of a distinguishable dimension of colour with a certain characteristic of light. The fact is that *any* particular phenomenal hue, including any of the unique hues, "can be evoked by *indefinitely many combinations of wavelengths*, none of which need have any wavelength in common with any other combination" (Hardin 1985, 44-45, emphasis added; see also Hurvich 1981, Chp. 4). No particular wavelength uniquely corresponds with a given hue. Monochromatic light of 503 nm, for example, will (for most subjects) produce a unique green colour experience. However, a mixture of 490 nm and 540 nm, or indefinitely many other wavelength pairs (none of which need contain 503 nm), will also produce unique green (Hardin 1984a, 495).

We are left, then, with an unsatisfactory situation. Different colours can not be identified with different wavelengths of light, nor can distinguishable dimensions of colour be identified with certain physical characteristics of light. We are without an explanation of why colour experiences can match and of the relationships between colours. But if the explanation we require is not to be located in the physical characteristics of light, where is it to be located? The answer, in short, is in the *organism*, as the following discussion of the processes subserving colour vision will reveal.

2.2. The Physiology of Chromatic Vision

The first task is to demonstrate why indefinitely many distinct combinations of wavelengths result in the same perceived colour, and explain how we are able to discriminate among wavelengths. To do so we have to return to the receptor system of the retina. When a receptor absorbs a photon, molecules of the photopigment of the receptor undergo a change in configuration (a process called "isomerization"). However, regardless of the wavelength of the photon which produces the isomerization, the change in shape is the same. Once a receptor absorbs a photon, the neural signal it produces is independent of the wavelength of that photon. This is known as the *principle of univariance* (Boynton 1979, 110; Hardin 1988, 26). If a receptor, upon presentation of light stimuli at different times, has the same number of molecules undergo isomerization, then it will respond in the same way regardless of differences in the wavelengths of the stimuli. Wavelength information, then, is irretrievably lost at a very early stage of visual processing when we confine attention to the output of a single receptor. Moreover, a single receptor cannot differentiate between wavelength and intensity. The output of the receptor is determined by the number of quanta absorbed. This is a product of the probability of absorption of a given wavelength, and the number of photons at that wavelength. For example, 6667 quanta at a wavelength of 575 nm and 1000 quanta at a wavelength of 500 nm will result in a receptor absorbing the same number of quanta, and its output will be the same in both cases (Cornsweet 1970, 157). Indeed, the response profiles of each type of receptor unit indicate a general pattern of behaviour. Each such receptor will respond with a very high degree of reliability to a certain stimulus, but it will also respond, though less reliably, to slightly different stimuli. Receptors, then, should not be considered as simple detectors of the values of features which impact upon them (Austen Clark 1993, 32).

The capacity for wavelength discrimination, and the reason why distinct combinations of wavelengths produce the same perceived colour, lies with the fact that there are three types of receptors – the cones – subserving chromatic vision. Each type of cone contains a different photopigment. The photopigments, via differences in their chemistry, differ in their ability to absorb light of different wavelengths. The three types of cones, in other words, differ in being maximally sensitive to light of different wavelengths. One type of cone has a peak absorption at approximately 430 nm, another type around 530 nm, and a third at about 560 nm. That is, we have short-wavelength, middle-wavelength, and long-wavelength receptors (hereafter often the S-system, the M-system, and the L-system). The sensitivity of each cone type at each wavelength is proportional to the probability that the cone type will absorb a quantum of light at that wavelength. So, while each cone type will be maximally sensitive to a certain wavelength, all three types of cone respond, in a weaker fashion, to light of wavelengths other than those to which they are optimally sensitive. Each cone type exhibits a sensitivity to a range of wavelengths, which can be expressed by its characteristic spectral absorption curve (see Hardin 1988, 27). Moreover, there is significant overlap in the range of wavelengths to which the cones respond. For example, light of wavelength 600 nm will elicit the strongest response from long-wavelength receptors (those which are maximally sensitive to light at 560 nm), but will also elicit some, though weaker, response from middle-wavelength and short-wavelength receptors. The degree of overlap is not, however, equal for the three cone types. The middle-wavelength and long-wavelength cones have spectral absorption curves which overlap to a far greater degree than the short- and middle-wavelength, and short- and long-wavelength, absorption curves (Hardin 1988, 26-31; Hubel 1988, 162-164).

It is the differing absorption profiles of the three cone types, and the overlap of these absorption profiles, which is the first step in an explanation of both differential responses to different wavelengths and the fact that indefinitely many combinations of wavelengths can result in the same perceived colour. Consider various pairs of single wavelength light, independently presented to a given cone type, say the middle-wavelength receptor unit. Certain pairs of wavelengths will result in the same percentage of photons (say 10 per cent) being absorbed by the M-system. This is because each of the three wavelength receptor systems is optimally sensitive to a certain wavelength, while still absorbing a certain proportion of incident photons at other wavelengths in its range of sensitivity. So, in the example, the M-system's output will be the same, despite differences in wavelength. The situation will be different, however, when we consider the response of one of the other cones, say, the long-wavelength receptor unit. Each one of a given pair of wavelengths which produced the same response in the M-system produces different effects in the L-system. So, for example, for one of the light sources, around 5 per cent of photons cause isomerizations, while at the other wavelength, about 20 per cent of photons cause isomerizations. If the outputs from the M and L systems are combined, the different wavelengths will produce differentiable effects. *Differences* in the relative outputs of the receptor systems are all that remain of wavelength differences in the light sources. Differences in receptoral outputs correspond to *ratios* of receptoral stimulations. A unique ratio between the outputs of the M and L systems is typically produced by any given wavelength. Normal colour perceivers (trichromats) have, as suggested, three types of cones. So, output differences are a function of three-term cone-excitation ratios. And it is in this "information" which the chromatic visual system trades (see Cornsweet 1970, Chp. 8; Hardin 1988, 26-33).

We can see, then, why the project of identifying any given colour with a particular wavelength of light is unlikely to succeed. And the first stage of an explanation of how our visual system is able to respond differently to different wavelengths, and how different combinations of wavelengths result in the same perceived colour, is in place. It is useful to pause and point out that it is in fact possible to *predict* which combinations of wavelengths will result in the same perceived colour for a given individual. If we know the absorption spectra for the receptor systems of a given individual, we can construct her *wavelength mixture space.*[103] Each axis of wavelength mixture space corresponds to one of the three wavelength systems (short, medium, and long), and so represents the number of absorptions of quanta in each system. Each point in wavelength mixture space, then, represents a certain number of absorptions in each system. And each point represents the effect in the three wavelength systems of indefinitely many combinations of wavelengths. The number of quanta absorbed by each wavelength system is, recall, a product of its characteristic absorption profile and the number of incident quanta. So while a given wavelength system absorbs the same number of quanta upon presentation of light of a given wavelength, the total number of quanta absorbed can be changed by changing the intensity of the light.

Wavelength mixture space is, then, a powerful tool for illustrating the principles underlying effects on, and thereby activation of, receptor systems produced by different stimuli. Once the absorption spectra for receptor systems of a given individual are known, that individual's wavelength mixture space can be constructed and predictions can be made about which combinations of wavelengths will result in the same perceived colour for the individual concerned.[104] Such predictions can be made by employing simple vector operations. The direction of a vector in wavelength mixture space represents wavelength composition; its length

---

[103] Wavelength mixture space is not to be confused with the "quality space" mentioned above. The relationship between the two will be discussed below.

[104] It's a long established fact of visual science that two stimuli which differ in terms of both their wavelengths and intensity can produce the same effect in wavelength systems. Cornsweet (1970, Chp. 8), for instance, provides a clear illustration of this point, showing how a single wavelength stimulus consisting of 1000 quanta at wavelength$_1$, and a wavelength-pair stimulus consisting of 1250 quanta incident at wavelength$_2$ plus 300 quanta at wavelength$_3$, produce the same effect upon the M and L systems (the S system is omitted to simplify the example).

represents stimulus intensity. The same effect on receptor systems by different wavelength combinations – that is, the same point in wavelength mixture space being reached – can be predicted by vector addition. Any given point in such a space can be reached by indefinitely many vector sums. Indefinitely many different combinations of wavelengths can result in the same colour – the same point in wavelength mixture space – and so the proposal to identify distinct colours with distinct wavelengths of light will not succeed (Cornsweet 1970, Chp. 8; see also Austen Clark 1993, 29-45).

The second step in an explanation of the chromatic system's response to different wavelengths and wavelength combinations – and, indeed, the second main step in a reductive explanation of the structure of qualitative similarities among colours – lies with the processes subserving receptor output comparisons. Here the centrepiece is the *opponent process* theory of colour vision. An opponent process can be thought of in terms of an opponent, or antagonistic, pair which engages in a sort of tug of war. Input to the process results in either some level of excitatory output, some level of inhibitory effect, or a neutral balance of excitatory and inhibitory activity (baseline activity). Opponent processes compare the ratios of cone outputs via different patterns of excitatory and inhibitory connections to the three cone types. There are three opponent processes: white-black, red-green, and yellow-blue. Each opponent process, or channel (with 'channel' understood as a functional notion), receives inputs from the three cone types (though the contribution of the S-system to the white-black channel, and the red-green channel, is very small). Activity in each opponent process is, however, independent of activity in the others. The outputs of opponent processes constitute neural signals sent from the retina to the visual cortex.

There is widespread acceptance of opponent process theory within the visual science community. Disputes focus not on chromatic opponency itself, but on precise mathematical characterisations of the theory, an issue I will not be concerned with here. A simple representative sketch of the central elements of the theory will serve my purposes. The sum of the outputs of the L and M systems generates the achromatic signal, the brightness channel. The difference between the outputs of the L and M systems (the neural counterpart of differencing is inhibition) produces one of the two opponent chromatic channels, the red-green opponent channel. The other opponent chromatic channel, the yellow-blue opponent channel, is generated by the output of the S-system being subtracted from the summed output of the L and M systems. Taking zero to stand for the base rate of neural activity, the neural code can be given, following Hardin (1988, 35), as follows:

(L + M) is the achromatic signal.

(L + M) greater than zero codes whiteness;

(L + M) less than zero codes blackness;

(L + M) codes for "brain grey", as do all zero values.

(L - M) is the red-green signal.

(L - M) greater than zero codes redness;

(L - M) less than zero codes greenness.

(L + M) - S is the yellow-blue signal.

(L + M) - S greater than zero codes yellowness;

(L + M) - S less than zero codes blueness.

The above formulation provides for brightness and hue, two of the three dimensions of perceived colour. The neural code for *saturation*, at a given wavelength, is the ratio of the chromatic output at that wavelength to the sum of chromatic and whiteness output at that wavelength. For each channel, a positive value means an increase in neural base rate firings, and a negative value a decrease in neural base rate firings. Which one of the pair of hues for each chromatic channel is assigned a positive value (for example, redness), and which is assigned a negative value (for example, blue) is, at current understanding, purely conventional. However, it is of course an essential part of the theory that redness and greenness, and yellowness and blueness, be given opposite signs (Hardin 1988, 34-35; 1984b, 126-127; see also Hurvich 1981, Chps. 11-13).

2.3. The Structure of Colour Phenomenology

The discussion of opponent process theory has been quite brief, though sufficient for the two main tasks to be undertaken below. The first is to show how the opponent process theory can explain the relationships between perceived colours. The second is to demonstrate how the theory underpins the structural isomorphism conception of the representation of colour. I have so far merely mentioned that colour experience exhibits a structure. It is time to expand upon that claim. I'll focus attention, for the moment, upon hue.

What does it mean to say that colour experience has a structure? In a series of publications, C. L. Hardin (1984b, 1985, 1988) has done much to address this issue. He is fond of beginning with a passage due to Armstrong (1978, 116):

> If we consider the class of shapes and the class of colours, then both classes exhibit the following interesting but puzzling characteristics which it would be agreeable to understand:
>
> (a) the members of the two classes have something in common (they are all shapes, they are all colours)
>
> (b) but while they have something in common, they differ in that very respect (they all differ as shapes, they all differ as colours)
>
> (c) they exhibit a resemblance-order based upon their intrinsic natures (*triangularity* is like *circularity*, *redness* is more like *orangeness* than *redness* is like *blueness*), where closeness of resemblance has a limit in identity
>
> (d) they form a set-of-incompatibles (the same particular cannot be simultaneously triangular and circular, or red and blue all over).

This passage dovetails nicely with the microcontent conception introduced above (Chp. Four, sec. 2.). The class of colours is a feature, dimension, or variable; all colours share, or exhibit, the property of being coloured. All colours differ in the very thing they have in common; they are all different values of a variable. The class of colours admits of a similarity structure in terms of the values of its members; a certain value is closer to some values than others. And, one member of the class cannot simultaneously be identical with two distinct values. A satisfactory answer to our question concerning the structure of colour would do well, then, to address Armstrong's *desiderata*. Indeed, in certain respects, we can be even more specific. Consider the following list of factors from Austen Clark (1993, 147-148):

> (1) There are just four unique hues (red, green, yellow, and blue). All the other hues are perceptually composite, and can be analysed into a particular ratio of two of the unique hues.
>
> (2) The unique hues comprise two opposing pairs. That is, red is least similar to green; yellow is least similar to blue.
>
> (3) Some yellow and blue lights can be mixed to result in an achromatic perception.
>
> (4) There are bluish greens, but no reddish greens.

(5) There is an orange which has equal components of red and yellow. That is, it is midway between red and yellow.

(6) Pink is more similar to red than it is to green.

(7) The hues can be arranged in a circle.

That there are four unique hues is well documented. There is also evidence that hue space is divided into four similarity classes based on the four unique hues, and that the unique hues thereby play a primary role in the relationships between colours.

In an experimental procedure reported by Hardin (1988, 42), subjects are shown randomised sequences of, say, 25 perceptually distinct monochromatic lights. Subjects are instructed to treat any presented light as consisting of a colour sensation equal to a value of 100 per cent. They are given a small set of colour names and requested to assign percentages to each of the colour names that best describes the amount of each of these colours perceived in each presented light. The subjects are told that if the colour names supplied do not completely capture any particular light, they are permitted to assign percentages to the colour names which do not add up to 100 per cent. The results of this experimental procedure lend considerable support to the view that the unique hues are a crucial determinant of the structure of hue space. In the version of the procedure where the set of colour names consisted of all, and only, four unique-hue names, subjects were able to completely describe all the presented lights. This was not so in the version where the set of names lacked one of the unique-hue names.

Further support for the primary role of the unique hues, and, indeed, an illustration of the relationships among hues, comes from the application of multidimensional scaling techniques (in particular, a variant called "proximity analysis") to a sample of 14 hues (of approximately equal brightness). Subjects were presented with two hues at a time and asked to rate the similarity of each pair on a scale from 1 to 5. When qualitative similarity ratings were subjected to proximity analysis, the result was the structure given in Figure 5.2. The fact that the unique hues are approximately ninety degrees apart, and that yellow and blue, and red and green, are approximately opposite one another is particularly illuminating, since this is just what opponent process theory would predict. The opponent process theory, recall, posits three opponent processes: red-green, yellow-blue, and white-black. Since the white-black process determines brightness, opponent process

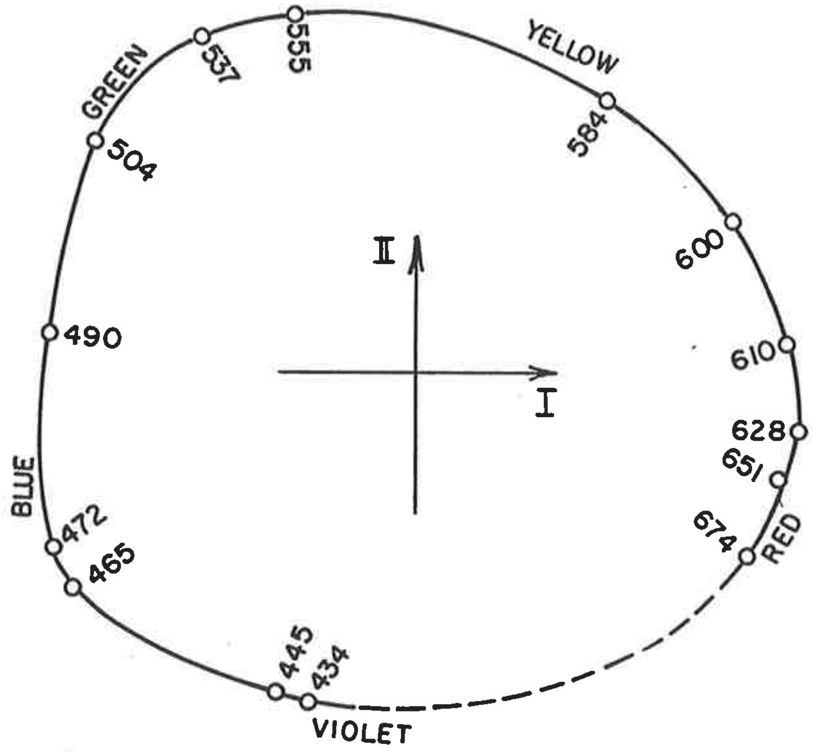

Figure 5.2. Proximity analysis of the ratings of the "qualitative similarity" of fourteen colours by adult subjects. (From Hardin 1988, 43).

theory can easily explain why there are just four unique hues, and why, for instance, there is a unique red but no unique orange. Opponent pairs are engaged in an excitatory-inhibitory "conflict". The unique hues occur when one of the chromatic opponent processes is in neither an excitatory nor inhibitory state – that is, it is responding at baseline – and the other chromatic opponent process is responding at either below or above baseline. So, for example, when the yellow-blue system is in a neutral state, it is solely the state of the red-green system which determines perceived hue. If the red-green system is excited, unique red is the result; if it is inhibited, unique green occurs. Similarly, when the red-green system is responding at baseline, the state of the yellow-blue alone determines perceived hue. If the yellow-blue system is responding above baseline, unique yellow is the result; if it is responding below baseline, unique blue occurs. This shows why there can be only four unique hues, and why the unique hues comprise two opposing pairs. There cannot be a unique orange, for example, because orange results from wavelengths in the spectrum which stimulate both the red-green, and yellow-blue, system. We can also see why orange is between red and yellow. Orange is produced by a level of activity above baseline in both the red-green, and yellow-blue, system. Orange is thus one of the binary hues. Every orange is to some degree reddish and to some degree yellowish. There is also a so-called "balanced" orange; that is, an orange which has equal components of red and yellow.

Opponent process theory can also explain why there are bluish greens, but no reddish greens. The many bluish greens (or greenish blues) result from various levels of activity below baseline in both the yellow-blue and the red-green system. Because red and green comprise one opponent system there cannot be a reddish green, since this would require activity in the system both above and below its baseline simultaneously, which is physically impossible.

Opponent process theory explains the existence of the unique hues, and the fact that the set of unique hues consists of two opposing pairs. The unique hues structure hue space into four similarity classes. The intermediate hues – the oranges, the bluish greens, the purples – result from various levels of activity in both chromatic opponent systems. The degree of qualitative similarity between hues is reflected in their arrangement in a hue circle.

Why is it possible to arrange the hues in a hue *circle*? Such an arrangement of qualitative similarity doesn't carry over to other modalities. For example, it is not possible to order pitches in terms of qualitative similarity in a pitch circle. The answer lies with two factors: the number of receptor systems, and their degree of overlap (see Austen Clark 1993, 153-157). There are only three types of receptors involved in chromatic vision (the cones). In contrast, for almost every audible frequency there is a receptor optimally sensitive to that frequency. However, the response profiles of *every* auditory receptor overlap with those of its neighbouring receptors. That is, there is not an auditory stimulus that affects only one type of auditory receptor. The situation is different for the chromatic receptors. The shortest and longest wavelengths to which the chromatic system is sensitive produce no response in at least one of the three receptor systems. The shortest wavelengths to which the chromatic system is sensitive produce activity only in the

S-system; the M-system and the L-system are unaffected. Most of the longest wavelengths to which the chromatic system responds produce activity in both the M and L systems, but not in the S-system. This means it is possible to "take the longest wavelengths to which any receptors are sensitive and mix in gradually increasing proportions of those affecting just the S-system. We can get a smooth progression of intermediaries from 100% L, 0% S to 0% L, 100% S. The two endpoints of the spectrum connect, and we have a hue circle" (Austen Clark 1993, 154).

Certain combinations of wavelengths result in achromatic perceptions. As suggested above, two lights which if presented independently would result in a yellow, and a blue, perception, can be mixed to yield a shade of grey. Such combinations produce a neutral state in both chromatic systems. The particular intensity of the grey perceived is determined by the white-black system. We also have to invoke the white-black system to explain pink. Pink results from activity above baseline in both the red-green system and the white-black system. In effect, pink is white "stained" with red (see Hardin 1985, 37-40, 1988, 36-45; Austen Clark 1993, 149-151).

I have explained, then, each one of the factors from the list describing aspects of the similarity structure within the class of colours. Let me now bring saturation back into the picture, and return to Armstrong's *desiderata.* Consider some colours which differ in hue and saturation, yet have the same brightness. Such a set is modelled in Figure 5.3. Saturation, recall, is the strength, or purity, of a hue; that is, it is inversely related to the achromatic, or grey, content of a colour. The centre of the circle in Figure 5.3. represents zero saturation (grey). Distances from the centre correspond to degrees of saturation; saturation increases from zero at the centre to 100 per cent at the circumference. The radii are lines of constant hue; all the hues are represented by angles around the centre. We have a constant-brightness hue-saturation space. Now, as Hardin (1985, 35-37, 1988, 113-116) has demonstrated, such a space can be employed to satisfy Armstrong's desiderata. With brightness constant at a certain level, every colour can be represented by a vector extending from the centre of hue-saturation space. Saturation is represented by the length of the vector, hue is represented by its direction. Each vector "can be resolved into exactly two orthogonal components which lie along unique-hue vectors; in the case of the unique-hue vectors themselves, one of the components is always zero" (Hardin 1985, 37). The relationships among the colours can thereby be mapped into a set of relationships among vectors, "reducing questions about an apparently qualitatively heterogeneous collection of objects to questions about a qualitatively homogeneous set of objects" (Hardin 1988, 114). This will satisfy Armstrong's *desiderata*. As Hardin writes:

> Thus, the members of an equal-brightness colour class
> (a) are all representable by vectors in the same H-S space;
> (b) differ in their vectoral components;
> (c) exhibit a resemblance order modelled by the angular separations and lengths of their vectors,
> (d) form a set of incompatible, in that no two distinct vectors can have both of their components in common.
>
> So the relations that the colours bear to each other are neither more nor less puzzling than the relations borne to each other by a set of vectors extending from a common origin in a space with four preferred directions (1988, 116).

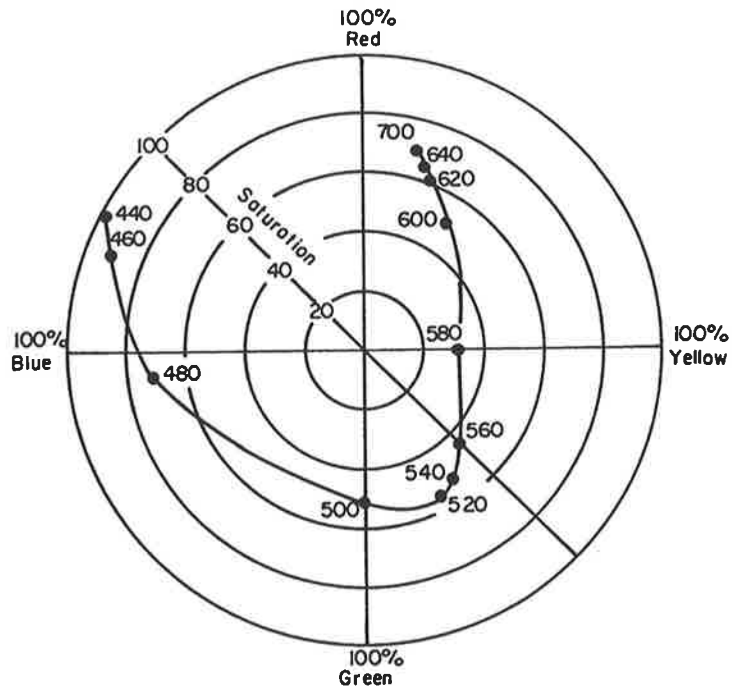

Figure 5.3. Hue-saturation space. (From Hardin 1988, 115).

The discussion of the similarities among colours moves toward its final destination by taking aboard *brightness*, the third dimension of perceived colour. The addition of brightness means the structure of colour experience can be expressed by a three-dimensional *quality space*: this is the Hue-Brightness-Saturation space (Hardin 1988), or psychological colour solid (Austen Clark 1993). Colour quality space is constructed by illustrating the range of achromatic values from dazzling (white) to dim (black) with a line orthogonal to the equal-brightness plane and passing through its centre. An equal-brightness plane, or hue circle, is, as discussed above, a roughly circular two-dimensional arrangement which is mapped using a polar coordinate system. One coordinate is hue, and the hues vary with angles around the circle. The other coordinate is saturation, which is the distance from the achromatic centre of the hue circle. Parallel equal-brightness planes through each value of the achromatic dimension yield a colour quality space (see Hardin 1988, 116; Austen Clark 1993, 119-120). A representative example is provided in Figure 5.4.

The dimensions of a quality space correspond to the differing respects in terms of which the members of a quality class are phenomenally related. Colour quality space has three dimensions since if either hue, brightness, or saturation differ in one experience to the next, the experiences are qualitatively distinguishable, and so belong to a different experience type. The coordinates of a point in colour quality space correspond to particular instances of each dimension, or attribute, of colour: a particular level of brightness, a specific hue, and a certain degree of saturation. Each point in the space corresponds to a different type of colour experience, where at least one of the coordinates differs from the coordinates of all other points in the space. A region in quality space specifies experiences which bear a close resemblance. A unique blue of medium saturation and brightness, for example, will be at the midpoint of a volume constituted by relatively small, but distinguishable, variations in hue, saturation, and brightness. The greater the distance between two points or regions, the less the corresponding experiences resemble one anotlner.

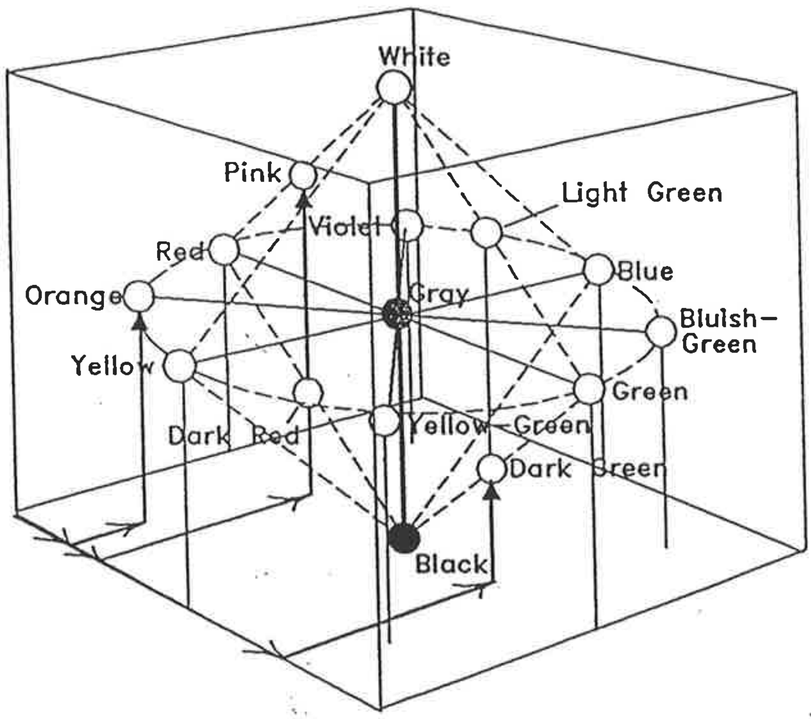

Figure 5.4. Colour quality space. (From Churchland 1995, 25).

2.4. The Neural Representation of Colour

As with the constant-brightness hue-saturation space, however, colour quality space provides no more than an illustration, albeit an extremely useful one, of the structure intrinsic to colour experience. Colour quality space is an illustration of a representing system under a phenomenological interpretation, a system of colour experiences. Sometimes a quality space is offered as a theory of content determination for a certain represented feature. But a quality space is not a theory; a quality space itself requires explanation. What is required is an *interpretation* of colour space. This is achieved, in part, by providing a non-phenomenal interpretation of the axes. As it stands so far, the dimensions of the representing system have been characterised only in phenomenal terms. My above claim (subsec. 2.1.) that the explanation of the structure within the colour representing system lies with the organism, then, is yet to be completely substantiated. A significant part of the required explanation is, however, in place. I have demonstrated above that opponent process theory goes a long way in accounting for the qualitative similarity structure within the class of colours. I will elaborate upon the case for an internalist explanation by investigating another attempt at an externalist account.

The above discussion has demonstrated the difficulties encountered in finding the required explanation in the physical characteristics of light. I can now also reveal that the explanation is not to be found in the effects of light on the organism's receptor, or transducer, system. The effects of light on the transducer system were illustrated above in terms of wavelength mixture space. Wavelength mixture space, recall, captures the effects of light stimuli upon the three wavelength systems. Each axis of wavelength mixture space corresponds to one of the three wavelength systems, and each point in the space represents the number of absorptions in each system. However, colour space can't be explained by the various possible *mixtures* of wavelengths. Here I signal a *crucial* point. Effects on the transducer system do not explain colour quality space because *colour space is not a constant function of wavelength mixture space*. Qualitative similarity between hues (to confine attention to an equal-brightness case) does not map in a constant way on to their location in wavelength mixture space. Pairs of wavelengths which are, for example, 10 nm apart are not all qualitatively similar to the same extent. The reason is that each opponent system takes as inputs the output signals from two and sometimes three receptor systems, but the outputs are not treated equally. That is, each opponent system accords a *different weighting* to the outputs of the three receptor systems. For example, excitatory outputs from the L-system result in a greater excitatory effect on the red-green system than the yellow-blue system (Austen Clark 1993, 151; Hurvich 1981, 133). This is due to weighted pattern of excitatory and inhibitory connectivity between the receptor systems and opponent systems. An explanation of the structure within the colour representing system, under a phenomenological description, is not to be found in the effects upon an organism's transducer system. So, colours can't be identified with, and so cannot be explained by, either *wavelengths* or *combinations of wavelengths*. Attempts to explain colour in an externalist fashion are not successful.[105]

To see the significance of the above point, consider a proposal to explain the character and content of colour experience offered by Paul Churchland (1988, 1989a). According to Churchland, the required explanation is to be found in the brain's coding of the reflectance properties possessed by surfaces in the external world. See Figure 5.5., and contemplate the following passage.

> Consider...the abstract three-dimensional "colour cube" proposed by Edwin Land, within which every one of the humanly discriminable colours occupies a unique position or small volume.... Each axis represents the eye/brain's reconstruction of the seen object at one of the three wavelengths to which our cones are selectively responsive (1989a, 102).

[105] Here I want to briefly return to my favoured reply to the "indeterminacy objection" levelled at the structural isomorphism TCD (Chp. Four, sec. 3.). The reply I recommended, to reiterate, is that isomorphisms are just not, as the indeterminacy objection would have it, ubiquitous. An analogue embodied in a mental representing system of a cognitive creature will be a very sophisticated structure, making it highly improbable that any given representing system will be structurally isomorphic to a range of disparate represented domains. I've just revealed that there is not a neat mapping of wavelength mixture space onto colour quality space, and thereby (by hypothesis), onto activation space (the representing system under a neurophysiological description). There is no structural isomorphism between the colour representing system and the system which receives external stimuli as input (and which passes activity onto the colour representing system). Here we have the *absence* of a structural isomorphism between two systems involved, at varying stages, in the *one representational task*. Given the lack of an isomorphism between two systems involved in our chromatic representational capacity, it seems highly unlikely that the particular structure of our colour representing system will exhibit an isomorphism to a number of disparate features of the representable world. And the way is open for the structural isomorphism theorist to claim that this result will generalise to other representing systems.

Part of Churchland's account appeals to the activity in the brain's colour representing system. The visual cortex receives the outputs of the three cone systems, resulting in a characteristic activation pattern (expressible as a three-value vector) for each distinguishable colour.

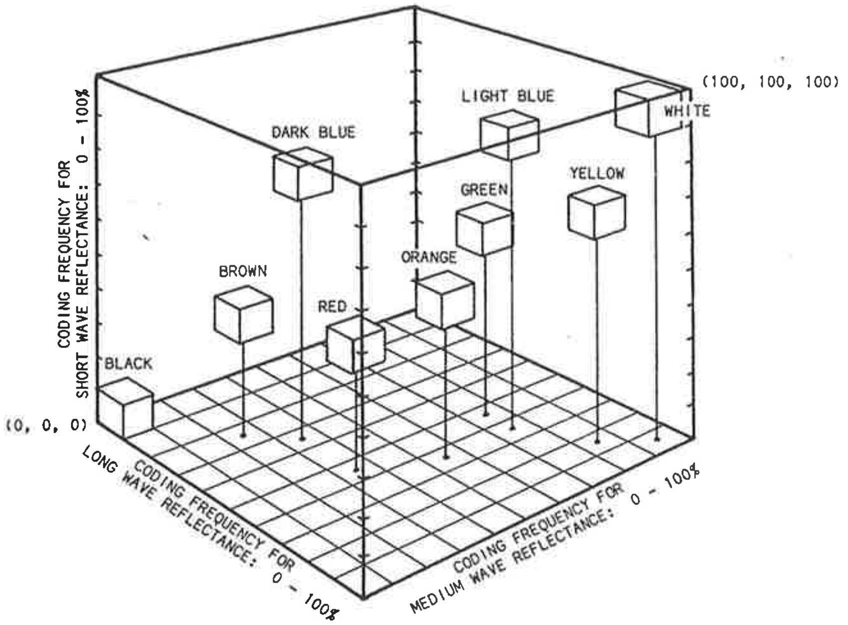

Figure 5.5. Churchland's "Color Qualia Space". (From Churchland 1988, 149).

Churchland's proposal, however, has both internal and empirical difficulties. First, consider the descriptions Churchland provides for the axes of the space and the points, or regions, within the space. The descriptions of the regions (the colour words 'red', 'orange', 'yellow', etc.) are not consistent with the descriptions of the axes (see also Fodor and Lepore 1996, 154). The description of the *axes* means that the points in the space correspond to physical properties of visual stimuli. That is, each point in the space is a three-term expression with each term corresponding to a value of short, medium, and long wave reflectance. The description of the axes does not yield a *quality* space, but a psychophysical space. The description of the *points* in the space, however, suggests that Churchland is assuming it is a quality space, a representing system under a phenomenological description. But in order to be a colour quality space, the dimensions of the space must be provided with a phenomenal description, such that the points stand for the qualitative character of different types of colour experiences. There is, then, a troublesome ambiguity in Churchland's employment of the "colour cube", prompting Fodor and Lepore (1996, 155) to write that Churchland "hasn't decided what his state space theory is a theory of."

Alternatively, if Churchland's colour cube is understood as it stands – if, that is, we grant the differential descriptions of the axes of the space and the points within it – the difficulties for his proposal are two-fold. On the one hand, the colour terms assigned to the points in the space express Land's theory concerning how the qualitative characters of colour experiences vary as a function of the psychophysical properties of light. And this lands Churchland with the difficulty raised above of trying to explain the phenomenology of colour experience by appeal to the effects of wavelength combinations on the organism's receptor system. The difficulty, as demonstrated, is that colour quality space is not a constant function of wavelength mixture space. Colour experiences can't be explained by the psychophysical properties of light because the effects of light stimuli on the organism's transducer system do not map straightforwardly onto the system of colour experiences. Churchland gets it wrong precisely because he is trying to preserve the standard externalist account.

On the other hand, if one attempts to interpret Churchland in accordance with his aim to explain the "eye/brain's" coding for the reflectance properties of objects in the external world, it is very difficult to see how Churchland's proposal implicates a system of *representation bearers qua* neural or physical system. Churchland's account does not move beyond the cones, the transducer system of the retina. But this is not a representing system for the representational experiences of colour, because colour experiences are ultimately instantiated by patterns of neural activity at the primary visual cortex. The visual cortex receives the outputs of opponent process mechanisms which *transform* the inputs they receive from the cone system. By remaining with the receptor system, Churchland not only encounters the above difficulty of failing to explain the structure of colour phenomenology, he is not offering a *TCD* for the contents of colour representational experiences.

The way out of these difficulties involves two steps. First, it should not be expected that a single state space will provide the whole story concerning the representation of a certain (consciously apprehended) feature. Rather, there should be a clear distinction between a representing system *qua* system of mental states under a *phenomenological* description, and a representing system *qua* system of mental states under a *neurophysiological* description. Second, chromatic vision must be understood in terms of opponent process theory rather than Land's Retinex theory.[106] Consider colour quality space and the following question from Fodor and Lepore:

> How does Churchland decide what gets represented by dimensions of state spaces and what gets represented by regions in the state spaces that the dimensions define? What decides, for example, that "brown" and "dark blue" correspond to regions rather than dimensions? (1996, 154)

Qua quality space, there is a straightforward answer to this question. It was suggested above that in order to constitute a quality space, the dimensions of a state space must be provided with a phenomenological description. And hue, saturation, and brightness are known to be the three dimensions of colour experience. Hue, saturation, and brightness are the phenomenal attributes in terms of which colour experiences are differentiated. A change in any one of these attributes results in a different colour experience. That hue, saturation, and brightness constitute the dimensions of colour quality space is, as has been demonstrated above, well

---

[106] Churchland starts off on the wrong foot by taking Land's Retinex Theory of colour vision as his starting point. The vast majority of those within the vision science community, in contrast, see Land's theory as having been superseded by opponent process theory. The main problem with Land's theory, in very general terms, is physiological evidence has revealed that in terms of chromatic vision the nervous system operates in a very different manner to that proposed by the Retinex theory. While some development of Land's theory continues, the current situation can be summed-up as follows. Everyone in the vision science community agrees that opponent mechanisms have a crucial role in chromatic vision; this is not the case concerning the mechanisms proposed by Land's theory. (See the Appendix in Hardin 1988, and Zeki 1985, for more on an evaluation of Land's theory as compared to opponent process theory). To be fair to Churchland, he has recently recognised that a sea-change in opinion has been underway for some time in the vision science community. In a more recent work (1995), Churchland's brief discussion of colour vision appeals to opponent process theory.

supported by both behavioural evidence from certain experimental procedures and neurophysiological evidence in terms of opponent process theory.

It was pointed out above that colour quality space is, as a representing system under a phenomenological description, in need of interpretation; that is, it requires a nonphenomenal explanation. Providing a nonphenomenal explanation of colour quality space consists of finding a neurophysiological interpretation. In short, such an interpretation is given by positing a physically structured system of neural representation bearers as the representing system (the representing system, that is, under a neurophysiological description). Consider the following passage from Austen Clark in which he sets-out what it means to provide an explanation of a quality space:

> To provide a neurophysiological interpretation for a quality space is to identify neurophysiological states and processes that stand in the *same pattern of relations* as those occupied by the points in the quality space. Quality space can be treated as a structural description. It says of quale X, whatever quale X may be, that it is more similar to Y than to Z, less similar to W than to Y, and so on. It says the same sort of thing about W, Y, and Z. One finds a neurophysiological interpretation for this structure when one finds neural states and processes a, b, c that instantiate those variables X, Y, Z. Some neural states and processes stand in just the same pattern of relations as those obtaining among points in the quality space. One thereby gives a neurophysiological explanation for the structure of qualitative similarities and differences (1993, 148; emphasis added).

Clark never once mentions structural isomorphism, yet the parallel between the above passage and the structural isomorphism conception of a representing system is clear. His "interpretation" of quality space is, in terms of structural isomorphism, the postulation of a physically structured, neural representing system. The same "pattern of relations" exists among the values in the representing system under both a neurophysiological and a phenomenological description. The multidimensional space, as indicated, needs to be given two different descriptions; in effect providing not one, but two, multidimensional spaces. *Qua* representing system under a phenomenological description, the space illustrates phenomenal structure, the structure of qualitative character. Hue, saturation, and brightness are its dimensions. *Qua* representing system under a *neurophysiological* description, the space illustrates neurophysiological structure, the values of a physical property of neural representation bearers. Opponent process theory provides a neurophysiological interpretation of quality space. The activity produced by the three opponent systems provides the dimensions of the (neural) representing system space.

To consolidate, then, the claim I'm arguing for here is that an explanation of the structure of colour phenomenology is to be found not in the external world, but in the physically structured, neural representing system. The hypothesis of structural isomorphism, in concert with opponent process theory, provides a neurophysiological explanation of colour quality space. It is now known that certain types of neurons at the primary visual cortex are specialised for the representation of colour. And it is these neural systems which receive the outputs of opponent mechanisms. There is in fact evidence that an opponent-process mode of operation occurs not only in the retinal system, but continues further along the processing chain. It is, however, not at this stage known whether opponency is the manner of operation of those neural systems which ultimately realise colour experience. That is, it is not clear whether neural systems in the cortex operate in an opponent fashion or simply receive the outputs of earlier opponent mechanisms (Hardin 1988, 35). While this issue is of neuroscientific interest, it does not much matter to the view being offered here. According to the line of thinking I'm pursuing, the qualitative character of colour experience is realised by the activation levels of those neural systems whose activity is the result of the three opponent systems. A neural system is a network of neurons, or, more likely, a collection of networks, dedicated to the representation of a certain feature (such as colour). The activity of a single neuron is its firing frequency. The various firing frequencies of each neuron in a network constitutes, at that time, the network's *pattern* of activation. The neural bearers of colour experiences are patterns of activation across networks dedicated to receiving activation from the opponent systems. The range of possible patterns constitutes the brain's representing system for colour.[107]

Such a system of patterns exhibits a similarity structure. The patterns which realise, for example, all the green experiences – that is, the set of green experiences which have distinguishable qualitative character in terms of small differences in hue, saturation, and brightness – will bear a close similarity. Such a

---

[107] Again I'm adopting a connectionist perspective on the brain's representing systems. I'll have much more to say on connectionism in the following chapter.

set of patterns will be least like the set of patterns which realise all the red experiences. Likewise, the patterns which realise the yellow experiences will be very similar, and least like the set of patterns which realise the blue experiences. Take, for another example, a class of colour experiences within a relatively narrow range of variation in terms of brightness, and let us consider the set of patterns which includes unique green, unique blue, and all the bluish greens in between the two unique hues. According to the view adopted here, if we were to observe the patterns which realise the distinguishable qualitative characters of such a set, we'd see a gradual change in the "shape" or profile of the patterns as we move from unique green through the bluish greens to unique blue. The patterns which realise those qualitative characters very similar to unique green would be very similar in shape; likewise for the patterns which realise qualitative characters close to unique blue. As we moved further from unique green into the greens with some blue component, we'd observe a systematic alteration in the shape of the patterns such that they would come to look less like the unique green pattern. Moving further into the bluish-greens to those with a greater component of blue than green, the shape of the patterns would come to resemble the shape of the unique blue pattern to a greater degree. Were we to compare the shape of the balanced blue-green pattern (that is, the pattern for a blue-green with equal components of blue and green) with the shapes of both the unique green pattern and the unique blue pattern, we'd observe a shape midway on the transition from unique green to unique blue. That is, it would exhibit the same degree of similarity to the shape of the unique green pattern as the unique blue pattern.

The transition from unique green through to unique blue would reveal a gradual and systematic alteration in the entire pattern; that is, its overall shape would change. A plausible speculation, however, would be that as we proceed (around the hue circle) from one particular unique hue to another (such as from unique green to unique blue) a certain "component", or better perhaps, "area", of the pattern would alter in shape more significantly than other areas. That is, a particular unique hue to unique hue transition would exhibit a *characteristic* change in the shape of the activation pattern. A transition from unique green through the yellowish-greens to unique yellow, for example, would reveal a pattern which changed in shape more so in a different "area" of the pattern compared to the unique green to unique blue transition. Indeed, as suggested, once we reached unique yellow, the shape of the pattern would be less like the shape of the unique blue pattern than any other pattern in the transition. A story with the same plot (with only the names changed) could be told for the patterns which realise qualitative characters from unique yellow through unique red and back to unique blue. Indeed, the story could be told in terms of any or all of the distinguishable dimensions of colour, but I expect that by now the reader knows how it is going to end.[108]

2.5. The Colour Notional World

> Is the object really green? really brown? Does it have any single 'real' colour at all? ... People who spend much time considering these cases have been known to give up the notion of true colour entirely. We once asked a scientist who performs research on colour vision why people think that most opaque objects have a real colour. His answer was, 'They do? How odd' (Boghossian and Velleman 1989, 103).

What I've attempted to do in this section so far is marshal some support for the kind of representing systems posited by the structural isomorphism TCD. A central concern has been to show that current (and widely accepted) scientific understanding of processes subserving the generation of colour phenomenology –

---

[108] There is a subsidiary question I should touch upon. Why do all the possible activation patterns which constitute the brain's colour representing system realise experiences with the qualitative character particular to colour? Different types of colour experiences differ in qualitative character, yet they all share a particular quality which is not possessed by other representing systems. There is a natural extension of the theory of structural isomorphism to address this issue of inter-system qualitative difference. The structural isomorphism conception of content determination relies on physical features of the representing system. It suggests qualitative similarities and differences *within* a certain domain are accounted for in terms of similarities and differences in the representing system's activation patterns. One possible answer, then, seems compatible with, perhaps even demanded by, the theory. In short, *it's all in the patterns*. That is, the qualitative character particular to various representational experiences arises from the character of the activation patterns which constitute the corresponding representing system. The patterns which constitute the brain's colour representing system have a nature which differs from that of the patterns making up the brain's (say) taste representing system. While each pattern within the colour representing system differs either slightly or significantly from every other pattern in the system, the claim is that all such patterns are in some sense more like one another in character than any one pattern is like a pattern of the taste representing system. An answer along such lines requires far more elaboration than I have space for here. It requires a healthy dose of connectionism, in particular the idea of how different patterns of weighted connectivity in different PDP networks give rise to differentially shaped hyper-dimensional activation spaces, and thereby systems of activation patterns with disparate natures. And, of course, we'd require an account of what makes certain representational contents the contents of phenomenal consciousness. For a lot more on the latter issue and a little more on the former see O'Brien and Opie (1999a, 1999b).

in the form of opponent process theory – provides the empirical underpinning for an explanation of colour experience in terms of a physically structured, neural representing system. The idea that colour phenomenology admits of a structure has been substantiated, and it has been argued that such structure can be explained by wedding the opponent process account to the structural isomorphism theory concerning the neural generation of colour phenomenology. The claim of the structural isomorphism conception of the colour representing system is that the similarity structure within the range of possible colour experiences is to be explained in terms of the similarity structure within the range of possible activation patterns which realise their qualitative characters and constitute the brain's colour representing system. I have also tried to demonstrate the futility of attempting to identify distinct colours with distinct wavelengths of light (or of identifying distinguishable dimensions of colour with different physical characteristics of light), and of seeking to explain colour solely in terms of the effects of combinations of wavelengths on the retina's cone systems. We have, then, an empirically plausible illustration of structural isomorphism's conception of a representing system.

It's time now to address the represent*ed* world of colour. If, as argued, an explanation of colour is not to be had in straightforward properties of the objective world but in the nature of the processes subserving an organism's experience of colour, then colours can not be identified with objective properties of the world itself. And so representational experiences of colour do not stand-in for straightforward physical properties of the external world. What, then, do representational experiences of colour stand-in for? The answer I will develop and defend in this subsection is that colour experiences represent a colour *notional* world.

One result which mitigates in favour of the idea of a colour notional world has already put forward. I revealed above (subsec. 2.4) the finding that colour quality space is not a constant function of wavelength mixture space. This shows that the generation of colour cannot be explained by effects on the organism's transducer system, and so colours cannot be identified with *combinations* of wavelengths. This is a highly significant result. It breaks the connection between the internal neurophysiological story concerning the brain's system of colour representation bearers and a *straightforward* feature of the world itself. Colour representation bearers do not stand-in for variations along a simple or elemental feature of the external world. As I will show below, the colour notional world is an *averaged, contextualised, perspectival,* "slice" across a range of disparate but interacting physical properties of the world. The further evidence for and illustration of the colour notional world provided below largely draws on a telling discussion due to Hardin (1990), though Hardin himself does not invoke the idea of a notional world. The following material also touches upon other, arguably more sophisticated, versions of an externalist account of colour.

In overview, Hardin (1990) argues against three attempts to reduce colour to some objective feature(s) of the external world: wavelengths of light (for example, Armstrong 1968); dispositions of objects to affect normal perceivers in standard conditions (as advocated by Smart 1975); and spectral reflectances (for example, Hilbert 1987). And a central argumentative theme is to question the legitimacy of the notion of a colour *illusion*. These two, intertwined concerns are clearly of relevance here, for it's possible to accept my above line of argument that colour experience does not find an explanation in the external world, but maintain that colour *experience* isn't "real" colour. Real colour is to be identified with objective properties of the external world; any disparity between experienced colour and real colour is an instance of a colour illusion. I won't be concerned with all aspects of Hardin's argument. Rather, I'll focus on those phenomena and arguments most suggestive of the colour notional world. In effect, my postulation of a colour notional world is an attempt to fill the void left by Hardin's elimination of colour from elemental, objective properties of the world itself.

One phenomenon Hardin (1990, 556-557) cites is that of *"metamerism"*, discussed above (subsec. 2.1.), though not under such a description. Metamers are chromatic stimuli that are physically distinct but occasion the same colour sensation. I demonstrated above how metamers make trouble for an attempt to reduce colours to wavelengths of light. Indefinitely many distinct combinations of wavelengths can result in exactly the same colour experience, for example, unique yellow. To explain an experience of yellow we have to attend to the cone excitation *ratios*. However, "...the physicalist who would reduce real colours to wavelengths of light should be able to pick out the *real* colours on the basis of physical considerations [of wavelengths] alone" (Hardin 1990, 557). This, we've seen, can't be done. We can't explain yellow (for instance) without reference to the operating principles of the organism's mechanisms of chromatic vision. To take another example, white is typically a result of light stimuli made up of appropriate combinations of all the visible wavelengths. However it can also be elicited by many physically disparate sources, for example, the superimposing of just two monochromatic light sources, of which there are indefinitely many distinct pairs. And as Hardin (1990, 557) writes: "Which of these, according to the account of color that identifies colors with wavelengths of light, is "real" white, and which is just "apparent" white? And by virtue of what principles does one make such choices?"

Metamerism also serves to undermine the "standard-observer/conditions dispositionalist" thesis (Hardin 1990, 560-561). On this view, colour experiences themselves are not colours, but rather stand-in for complex conjunctions of physical properties resident in objects' "causal powers" (to employ terminology inspired by Locke). This thesis is an improvement over wavelength reductionism, for the character of a colour sensation typically depends (at least) on both surface reflectance properties and the nature of the illuminant. Now, metamers for a particular observer under a particular illuminant are two phenomenologically equivalent colour sensations elicited by spectrally distinct surfaces. The problem for the dispositionalist thesis here is that because the two surface stimuli are spectrally distinct, a *different illuminant* will result in different *sensations* for the two surfaces. It is well known that two stimuli which match well under daylight for the majority of the population elicit differing sensations under some kind of artificial light.

Of course, the dispositionalist will respond that it is the colour elicited by the standard illuminant which is the "true" colour. But the difficulty here is that the notion of the standard illuminant must now be specified, and any such specification does not look to be sufficiently principled. Sunlight and north daylight are two currently employed instances of "standard illumination", but their spectral profiles differ such that two colour samples will match under one illuminant and not the other. Again, on what principled grounds is one illuminant to be accorded the privileged status of *the* standard illuminant? Indeed, specifying the standard illuminant is only the beginning. The proponent of the "standard conditions/observer" thesis must formulate, in a non-arbitrary fashion, a standard for all those factors which are a determinant of perceived colour. These run to a considerable number. "Standard conditions" include not only the standard illuminant, but also: some standard of "neutral" *surrounds* (colour patches can look different with differently coloured surrounds, and some colours are even a product of such "simultaneous contrast", a phenomenon to which I'll return shortly); some standard of the *angular size of the colour sample in relation to the observer's eye*; and some standard of the *illuminant-sample-observer viewing angle* (Hardin 1990, 562-563). On the latter, viewing and illuminant angles influence perceived colour over a range of different kinds of materials. Some surfaces, for example, are glossy, which is an angle-dependent feature. Even more telling, much colour arises from such angle-dependent processes as scattering, refraction, polarisation, and interference (Hardin 1990, 563). As regards "angular size", there are two standards currently employed, but they yield slightly different colour phenomena, and colour technicians deploy one over the over solely on the basis of the task at hand. And the idea of a "neutral" surround must be understood in relation to the colour patch stimuli. There is no one neutral surround (Hardin 1990, 562).

So, again, a specification of "standard conditions" must incorporate all those characteristics and processes which can lead to a difference in colour phenomena. But because of the sheer number and diverse nature of such characteristics and processes, any specification is going to be arbitrary in regard to at least one or other. Any specification will inevitably fail to subsume some factor which makes a colour difference in some circumstance. It looks highly unlikely that a principled specification of "normal conditions" – no matter how highly complex a conjunction of factors it cites, and no matter how "gerrymandered" – will be forthcoming. The notion of the "standard observer" is similarly problematic. This reference point is "...actually a standardized set of color matching curves that are based on average values obtained from the color matches made by fifty or so normal – that is, non-color-deficient – observers." It is but an average, such that around 90 per cent of people would not completely agree with the colour matches of the standard observer (Hardin 1990, 563). The idea of "normal conditions/standard observer" is, then, a construct utilised for practical purposes by colour technicians in colour technology, one lacking the robustness upon which to hang a philosophical theory.

Consider now the well-known and pervasive phenomenon of *simultaneous contrast*. It is easily demonstrated with a few pieces of differently coloured paper.

The same piece of paper can lead to significantly different colour perceptions when surrounded by areas of differently coloured paper. In general terms, the phenomenon is that a large patch of colour is prone to incite its complementary colour in an adjacent colour patch. For example, a large patch of red renders neighbouring patches more green, blue induces more yellowness into an adjacent area, white makes surrounding regions look more black, and so on (Hardin 1990, 558). The notional world of colour is *contextualised.* (Incidentally, simultaneous contrast receives a straightforward explanation in opponent process theory. Indeed, anyone with the barest knowledge of opponent process theory could easily generate the correct explanation).

Again, the theorist who would identify colours with clear-cut properties of the external world will object that this is an anomaly, a colour illusion. The ubiquitousness of simultaneous contrast has such a claim ring hollow. Moreover, simultaneous contrast *must* be invoked to explain black and brown. These two colours, that is, are only ever produced by the principle involved in simultaneous contrast. Take brown as an example. The browns are just blackened oranges and yellows. Their dominant wavelengths coincide with those objects which standardly appear orange and yellow. The usual difference in perceived colour is solely a matter of the intensity of light reflected from a "brown" object and a "yellow" or "orange" object. A chocolate bar reflects light of a significantly lower intensity than an orange. That is, "[t]he characteristic difference in appearance between the two depends entirely upon their perceived relationships to the ambient light" (Hardin 1990, 559). If, once more, the externalist objects that these are merely aberrant cases, one must ask whether a satisfactorily complete account of colour could simply abdicate responsibility for black and brown. As Hardin notes: "Browns are, for most people, a distinctive set of colours, as differentiated in character from reds and yellows as reds and yellows are differentiated from one another" (1990, 559).

I turn now to the question of whether the "spectral reflectance" thesis can provide a satisfactory analysis of colour phenomena. The first difficulty, as Hardin (1990, 564-5) notes, is that the reflectivist must show how the theory can be extended to subsume those causes of colour which are not grounded in the reflection of light. As indicated above, these run to a significant number, and the required broadening of the theory is not a straightforward task. More significantly perhaps, a theory of colour must provide a satisfactory account not only of the hues of red and green, etc., but also of the structure inherent to colour experience, in particular the hue classes of red, green, yellow, and blue. The hues and their structure are an ineliminable fact of the phenomenon of colour. By identifying "true" colours with the spectral reflectances of surfaces in the external world, the reflectivist is without an account of the four-fold hue structure we find in colour sensation, that which, as indicated above, organises the structure of the whole range of colour experiences. As Hardin writes: "This fourfold color structure has no counterpart in physical structures outside the organism, and any attempt to assign reflectances to fourfold color classes will inevitably appeal to normal observers and standard conditions, with inevitable arbitrariness" (1990, 565).

Let me mention, then, just two further points. First, a well known finding of visual science is that the sensed colours of objects tend to be relatively impervious to rather significant changes in illumination. This is due to the fact that illumination changes are downplayed by mechanisms of chromatic vision. The chromatic visual system has been adapted to cope with illumination changes, such that the relative similarity structure of perceived colour is preserved. Tellingly, the relationships between the colours in the visual layout are more salient than the absolute values of the colours. The colour represented world is an "averaged" world. However, this phenomenon of "colour constancy", as it is known, should not be overstated (a charge Hardin makes of Land and subsequent advocates of his theory). Colours remain largely constant only within a certain range of illumination changes. Colour constancy begins to break down under certain natural lighting conditions, becomes fragile with artificial colourants and illuminants, and doesn't apply to changes in the *nature* of the illuminant (Hardin 1990, 561-562). Second, while the "settings" of certain external factors may not always have significant impact upon perceived colour, the "settings" of certain *internal* mechanisms play a notable role. To cite Hardin once more: "...the colors that we actually see depend upon many...factors..., such as the intensity with which the receptors are stimulated, what is going on in surrounding receptors at the moment, and what went on in the receptors during the previous milliseconds" (1990, 565). The colour represented world is contextualised and agent-centered.

In general, then, there is nothing in the world itself that straightforwardly lines-up with the structure of colour sensation. For this reason and others rehearsed above, externalist accounts of colour don't succeed. Moreover, the alleged distinction between "true" colour and "illusory" or "appearance" colour cannot be maintained. *All* the external world's colours are, in a sense, illusory. That the world is presented as coloured is, I claim, an illusion generated by the fact that colours – which receive their explanation in the chromatic system of the organism – are *"projected"* onto the external world by the intentionality (the "directedness", or "pointing outward") of our representational experiences of colour. And the represent*ed* world of colour does not coincide with the objective world; it is a purported or notional world. The colour notional world, as set out above, is an averaged, contextualised, agent-centered world. It is the world as-represented by an organism's colour representing systems. Such physically structured representing systems *impose* a structure upon the messy details of the world's disparate physical properties. They "carve" a structured represented system out of a range of diverse but interacting physical properties of the world itself. In this sense, then, the colour notional world is fundamentally a *"cleaned-up"* world.

3. The Notional World: A Projectivist Theory of Intentionality

In the previous two sections I've attempted to illustrate the idea of a notional world in regard to the representational realms of temperature and colour. The characterisations provided therein can be rehearsed to describe a notional world in general.[109] The move from particular examples to a hypothesis about representational content in general is, of course, fraught with difficulty. Any such extrapolation is, by nature, highly speculative. Nonetheless, if the arguments of the previous two sections are close to on-track, then the orthodox conception of content cannot be true of mental content in general. So, an extrapolation from the above particular examples is justifiable on the grounds of thinking about the world of mental content in more general terms. That is, a general replacement for the idea that mental content straightforwardly corresponds to the world itself is required, so the idea of a generalised notional world is worth broaching. Moreover, while a more complete defence of the notional world in general clearly requires a detailed

[109] The choice of an appropriate term for the reconception of the world of representational content I'm advocating herein has proven a difficult one. I've settled on "notional world" because it is the only expression in the literature which comes close to capturing the idea I wish to defend. The term is due to Dennett (1987, Chp. 5). At one point, Dennett describes the notional world as the "world-as-represented", contrasting it with the "real world" (1987, 172). I should hasten to add, however, that I don't use the term in exactly the same fashion as Dennett.

investigation into many domains of mental representational activity, a little reflection on representational capabilities less "peculiar" than colour (in particular) shows that it is worthwhile contemplating the idea of a generalised notional world. I assume, for example, that those representing systems which present the local spatial layout and general shapes of objects are (on most occasions) reasonably faithful reflections of the world itself, but even here the two do not straightforwardly coincide. The local spatial layout (and object shape) is represented as from the organism's point of view. And well known factors of visual perception to do with perspective, compression of distance the further away objects are from the perceiver, the limits of the focal point, and so on, show that the world itself is not simply laid out before our eyes in our visual representations. So, even the visual representation of the local environment – an indisputably *basic* representational capability – exhibits certain characteristics which conform to those I've portrayed as belonging to the idea of a notional world. I now propose, then, to sketch the concept of a notional world in general terms, and defend it against certain objections from the orthodox conception of content and content determination.

A notional world is the world as it purports to be by the internally generated content of an organism's repertoire of representational capacities. It is the world-as-represented by a creature's representing systems. It is thereby in some sense a "construct" of those representing systems, and determined by the number and kinds of representing systems a creature possesses, the dimensions by which such representing systems are circumscribed, and the resolution and skewing of the set of representation bearers which constitute each representing system. A notional world is a perspectival, contextualised, averaged, and, fundamentally, *cleaned-up* world skewed to the organism's constitution, capacities, environment, needs, and interests.

A notional world is, at once, the portrayal of *certain* properties that do have a reality beyond our representation of them, *together* with both certain mind-dependent features and a "construction" of the world as determined by the set of representing systems possessed by a kind of cognitive creature. This is by no means a surrender to idealism; there is an objective world. The brain's neural systems take a "slice" of the world itself. The particular slice taken is a product of the natures – the delineation and value-structures – of the representing systems involved. The question of what degree of "overlap" obtains between a notional world and the objective world is a difficult one that I don't propose to consider in this work. And it's yet another problem epistemologists may say they can very well do without. I want to tread, then, a delicately balanced middle path, asserting that there is some relationship between a cognitive creature's notional world and the world itself, while maintaining that a notional world – a world of representational content – is not equivalent to the world itself.

The reconception of the represented world and content determination being proposed herein will no doubt meet with resistance from those advocates of the orthodox view. I now propose to examine certain objections, and reasons for the predominance of the prevailing conception. One of the reasons for the strong grip exercised on the imagination by the orthodox conception is to be found, I think, in the following objection to an internalist theory of content determination. "The contents of perceptual experiences", our defender of the orthodox conception claims, "are to be *determined* by external objects, because it is external objects which cause perceptual experiences. That this is so can be demonstrated by attending to one's perceptual experience. Perceptual experience presents the external world. Examine the contents of your experience and you will find objects residing in the external world; what you will find, that is, is the world itself. So the claim that representational content is determined by factors internal to cognitive agents is mistaken". Let me break this objection into two parts. The first part – that content determination is externalist in nature – has been dealt with above. The mistake is a conflation of the tokening-feature and the content-feature. The alternative conception of content determination being proposed here does not deny that perceptual experiences are (standardly) tokened by external items. What is denied, however, is that content determination is to be thereby understood in terms of the tokening relation. While perceptual experiences are (typically) caused by an external feature, what is represented is determined in an internalist fashion, by, that is, the representing system itself.

The contrast between the internalist and orthodox conceptions of content and content determination is rendered particularly salient with the reply to the second part of the objection. The objection claims that intentional states, perceptual experiences in particular, present to us an external world. That is, the contents of certain intentional states are presented as residing in the external, mind-independent world. In the tokening of representational states, we are directly apprehending the external world. What is represented, the objection claims, is beyond a cognitive system, so content must be externally determined. The "connectedness" conception of intentionality must be true. The problem with this objection concerns a failure to recognise that there is another way to account for representational states presenting to us the external world.[110] Indeed, the appearance of a *direct connection* with the external world via representational states is, the reply urges, something of an illusion. (It is only "something" of an illusion because, of course, there is a mind-independent world, and it is the external world that very many of our representational states stand-in for). What I'm questioning here is the idea that only an externally initiated content-determining relation – like a causal relation – is capable of presenting an external world to cognitive subjects. The alternative, of course, is that intentionality be understood as "directedness". Representation bearers are pointed at a notional world, a world which purports to be the world as it is independent of its representation. Intentional states are directed outward. So, it is not that representation bearers – in being intentional – *connect* us to something external. Rather, representation bearers – in being intentional – are *directed* outward. Their intentionality gives the cognitive subject his presentation of a world beyond himself. On this view, internally generated content is a *projection* onto the external world. The brain's representing systems build a notional world which is a projection onto the external world.

This alternative conception of intentionality will strike many as counterintuitive, and, in the grip of the orthodox view, they will claim that it is just obvious that our perceptual states connect us with the external world, as even a casual inspection of one's conscious experience will demonstrate. But the projectivist conception is not as implausible as it initially appears. Granted it *seems* as if we are in *direct contact* with the world via our intentional states, *but this is because the fundamental characteristic of intentional states is that they are about something beyond themselves.* To be "in" an intentional state – to have an intentional state tokened – is to be presented with the content of that state as being something beyond that state, as something in the external world. And this is so not because we are in direct contact with the external world due to the intentional nature of mental states, but because we have an external world *with* intentional

[110] Recall also the conflation of a "theory of content" and a "theory of content determination" I raised above (Chp. Three, subsec. 1.5.).

states. Content determination, then, is not a matter of being causally connected to the external world, a matter of a causal impingement by the external world upon us. It is a matter of a *complementarity* of – a structural correspondence between – a purported world and internal representing systems. The content-determining relation doesn't hold because of, or isn't initiated by, external items, those which impinge upon our internal states. Rather, on the conception of intentionality and content determination being proposed here, it is only by virtue of something internal – a system of neural items having a systematic structure – that a correspondence relation, the content-determining relation, is generated. The content-determining relation is internally initiated and secured, and content is internally determined.

A further objection to the notional world and the "directedness" conception of intentionality – another intuitively powerful source of the "connectedness" conception – points to our not inconsiderable degree of behavioural success. We exhibit intelligent behaviour, that which is appropriate to our environments. According to advocates of the orthodox conception, the only explanation is that internal states connect or make contact with the world itself. In order to behave appropriately in the world itself it must be that the objective world exerts a strong influence over our behaviour, at least in the *de facto* sense that internal states represent the world itself and such externally determined content drives our behaviour. There is a faint whiff of a decaying behaviourism about this claim. What I'm proposing herein, in contrast, is a version of cognitivism which accords the crucial content determining role to neural representing systems rather than the causal impingement of the world itself. But this does not mean that it lacks the resources to explain behavioural success.

I have two lines of reply. First, and as indicated earlier, there is a degree of "overlap" between a creature's notional world and the world itself. The two are not wholly distinct in that they display no amount of convergence. Indeed, our behavioural success is taken by the advocate of the notional world to be some kind of *evidence* for such an overlap. Certain proprietary representing systems (or coalitions of such systems) will have notional represented features that overlap with the feature itself to a greater extent than others. As pointed out above, I think that those representing systems which represent the shapes and locations of objects in the local environment typically do so in a manner which largely conforms with the objective world, though, again as suggested above, the local environment "as-represented" doesn't exactly line-up with the world itself. Some representing systems are more "narcissistic" than others (Akins 1996). Recall the kind of "skewing" of temperature representation that occurs at certain extreme temperature ranges (sec. 1. of this chapter). This sort of exaggeration makes perfect sense from the perspective of the organism, in that those temperatures can *damage* the organism. This brings me to my second (related) reply.

There is a way of reading 'appropriate behaviour' (or 'behavioural success') as behaviour that is merely *in accordance with* the world itself, behaviour that is "faithful" to the objective world. Intelligent behaviour is behaviour that is directed upon the world itself which is appropriate to the objective world in that it is not "crazy" or perverse. Concomitantly, there is a way of interpreting the relationship between representation of the world and appropriate behaviour as the former being the slave of the latter. It is as if representational capacities evolved independently of those behaviours directed toward the represented features. The picture suggested here is that evolution designed creatures capable of intelligent behaviour by conferring upon them total and veridical models of an external reality. Equipped with such a rich, overall body of information about the objective world, creatures were thereby free to behave on the basis of such information in a manner appropriate to the objective world. Representing systems, then, are humble servants in the business of faithfully recording veridical information on the world itself to a high degree of fidelity. They are passive, unbiased, detached repositories of information. This view strikes me as somewhat under-described. Appropriate or intelligent behaviour is not just behaviour that is appropriate or faithful to the world itself. It must be understood relative to organism's constitution, behavioural repertoire, interests, needs, current state and proclivities, capacities, and so on. Appropriate behaviour has an agent-centered component. And a notional world, as indicated, builds-in such an agent-centered element.[111] Internally determined content – proprietary, physically structured representing systems which in some sense self-determine their represented systems – arose, at least in part, in the service of the "interests" of kinds of organisms.

[111] One must be careful, however, not to slide from this claim to the view that biological function, representation bearer "consumers", or cognitive role determines content.

Indeed, the projectivist theory of intentionality is not only a viable alternative to the orthodox view, it does, I think, somewhat paradoxically afford a more robust account of intentionality. Early on in this work (Chp. One, subsec. 2.1.) I distinguished between the question of representation and the question of content determination. The prevailing approach, however, is not to distinguish between these two questions. Instead, the naturalisation of intentionality project is seen solely in terms of the question of content determination. The reason for this situation is, I think, a commitment to the view (be it explicit or implicit) that representation bearers are symbols. The thinking seems to be that if we have a theory which attaches contents to symbols – a theory Cummins (1996, 66) calls the "game of *pin the meaning on the symbol*" – we will have automatically shown how the physical states of the brain which instantiate symbols possess intentionality. But the game of attaching contents to symbols assumes, rather than explains, intentionality. A theory of content determination must be embedded within a theory of representation. Attaching contents to symbols does not go to the question of how states of the brain are able to represent in the first place; it does not address the issue of how physical states could point beyond themselves to an external world. My claim here is that the "connectedness" conception of intentionality stops short of an account of the essence of intentionality.

Whether one's theory involves an explicit commitment to symbols (like Fodor's account), or says little on the nature of representation bearers (like Dretske's position), one thing is clear: the TsCD belonging to the orthodox approach view the intrinsic physical properties of the brain as largely irrelevant to content determination. There is an arbitrary relation between the intrinsic physical properties of a symbol type and its content. What matters is the tokening relation between the external world and internal representation bearers, and content is thereby identified with the representational object itself. On this view, the intrinsic physical properties of the brain's representation bearers make no contribution to the determination of content. There is merely a covariation of environmental object and the tokening of an internal representation bearer. It is only by virtue of such a covariation that it can be said that the representation bearer represents the external object. Representation bearers are mere detectors of external objects; they simply "light up" (Cummins 1996, 74) in the presence of a certain type of external object. But detectors don't present an external object to the organism. Mere covariation of detector "firings" with the presence of external objects does not give an external world to an organism. This is what goes unremarked in discussions of the main existing TsCD. The essence of intentionality – the pointing outward, the presentation of features to an organism as being beyond that organism – is left unexplained. It is not explained how the brain produces an external world.

It is the claim of the "directedness" conception of intentionality that the presentation of content to an organism is due to its internal determination. The world of representational content – the notional world – is constructed by the brain's representing systems. The notional world is presented as external because the intentionality possessed by the brain's representing systems is to be conceived of in terms of a projection. Intentional states do not put us in *direct* contact with – connect us with – the external world. Rather, intentional states are that by which we have an external world. In contrast to the orthodox understanding, what is being proposed here is a conception of intentionality which fully appreciates its remarkable nature, one which accords a crucial role to that vastly complex machine we have in our heads. We might *behave* in the world itself, but *how* we behave is governed by the internally determined world of representational content, the notional world.

CHAPTER SIX

COMPUTATION AND CONTENT DETERMINATION

In this final chapter I further develop and defend the structural isomorphism TCD by taking the discussion into the territory of cognitive science. The foundational claim of cognitive science is that cognitive processes consist in computational operations over neurally-instantiated mental representation bearers. As everyone knows, however, there are two competing computational accounts of cognition: the *classical* computational theory of mind (based on conventional computational theory), and the *connectionist* computational theory of mind (founded on the parallel distributed processing, or just PDP, computational framework).[112] My main undertaking will be to show how classicism and connectionism have differing implications for the question of content determination. In setting out to examine the relationship between each of the two computational frameworks and the issue of content determination, I embark upon a task which goes very much against orthodox thinking. It is standardly (if tacitly) assumed that the naturalisation of intentionality is orthogonal to cognitive science. While such an assumption is rarely explicitly endorsed, it is certainly manifest in the current practice of the naturalisation of intentionality project.[113] The prevailing strategy is to address the question of content determination in isolation from cognitive science in general, and the differing perspectives of classicism and connectionism in particular. In this chapter, I propose to challenge the view that the "orthogonality assumption" is true of cognitive science in general. I intend to demonstrate that while classicism is not forced to embrace any particular answer to the question of content determination, *connectionism* is committed to the structural isomorphism TCD.

The detailed line of argument is conducted over the course of section 1. Classicism and connectionism, it is widely assumed, are both *computational* accounts of cognition.[114] I spend some time attempting to make clear just what is

[112] Though I do spend some time detailing both the conventional and PDP accounts of the physical realisation of computational operations, I will be assuming a basic familiarity with classicism and connectionism.
[113] A notable exception is Cummins (1989, 1996).
[114] Certain theorists, however, reserve the term 'computation' for conventional computation, some clearly signaling this restriction (for example, Copeland 1993), and others not (for example, Van Gelder, 1995).

involved in the idea of computation, and arrive at a characterisation of computationalism about the mind neutral between classicism and connectionism; a characterisation, that is, of *generic* computationalism (subsec. 1.1.). Central to this characterisation is the notion of *semantic coherence.* Cognitive processes, as computational operations, consist in semantically coherent, or content-appropriate, transactions among mental representation bearers. Having arrived at a conception of computationalism which subsumes both classicism and connectionism, I then set about showing how the two computational theories of mind differ. I will examine both the conventional and PDP accounts of the physical realisation of computational operations (subsecs. 1.2. and 1.3., respectively). Classicism and connectionism, in being based on different computational frameworks, are standardly taken to constitute competing computational theories of cognition (though the exact nature of the difference is a matter of some controversy). According to classicism, cognitive processes are the operations of a collection of interconnected, neurally-realised conventional (that is, digital) computers. Cognitive processes are constituted by rule-governed, syntactic manipulations of mental symbols. The representation bearers of cognition are the simple and complex symbols in the language of thought. According to connectionism, cognitive processes are the operations of a host of interconnected, neurally-realised PDP-style networks. Cognitive processes consist in the spread and transformation of patterns of activation across neuron-like processing units linked together by differentially weighted connection lines. It is the patterns of activation and patterns of weighted connectivity which constitute the representation bearers of cognition.

It is at this point that the notion of semantic coherence becomes crucial. The causal operation of a computational system is *disciplined* such that representation bearers are processed in a semantically coherent fashion. I will suggest that a fundamental difference between classicism and connectionism lies with the contrasting ways in which the causal operation of conventional and PDP systems are disciplined. And it is this difference which relates to the issue of content determination. In classical cognitive science, semantic coherence is secured in a manner which is independent of an answer to the question of content determination. In connectionist cognitive science, on the other hand, semantic coherence must be understood as a product of the structural isomorphism TCD. The structural isomorphism TCD gives connectionists something they have long sought, but have yet to fully articulate: it shows how cognitive processes can be understood as something *other* than the operations of a conventional computational device while still consisting in *computational* operations.

1. Structural Isomorphism: A Theory of Content Determination For, and By, Connectionism

Over the past decade or so, cognitive science has become a very vibrant research programme. Ever since the advent of connectionism, an enormous amount has been written on the relative merits of the two competing computational frameworks. Moreover, certain theorists have recently begun to examine the brain's generation of phenomenal consciousness from the perspective of cognitive science, often (either implicitly or explicitly) aligning themselves with either classicism or connectionism.[115] Somewhat surprisingly, however, there has been a reluctance to examine the question of content determination from the perspective of cognitive science in general, and from the differing frameworks of classicism and connectionism in particular. Issues concerning the naturalisation of intentionality (in particular, the question of content determination) are typically thought of as orthogonal to issues within cognitive science (in particular, the nature of cognitive architecture). Even more surprisingly, one finds this attitude among committed cognitive scientists. One would think that since a commitment to cognitive science is a commitment to mental representation, we would not find cognitive scientists endorsing the view that the question of content determination is not a job for cognitive science. Curiously, however, this is not the current situation. The following passage from the opening chapter of Horgan and Tienson's recent book nicely captures the prevailing attitude:

> In this book, we have taken representation for granted. In doing so, we simply follow connectionist (and classicist) practice.... Every connectionist model has representations; otherwise it would not be a cognitive model. In both classical and connectionist modeling, content is assigned to certain states, essentially by the modeler's fiat.... It is assumed that natural cognitive systems have intentional states with

---

[115] Typically it is the connectionists who are forthright about their doctrinal allegiance. See, for example, Churchland (1995), Lloyd (1991, 1995, 1996), and O'Brien and Opie (1999a, 1999b). On the (often tacit) classical side, see, for example, Baars (1988), Dennett (1991), and Shallice (1988).

> content that is not derived from the interpretating activity of other intentional agents. But *it is not the business of either classical cognitive science or connectionist cognitive science to say where underived intentionality "comes from"*....(1996, 131; emphasis added).

I will argue that the prevailing view is a product of the *classical* computational theory of mind. The "orthogonality assumption", I claim, is *false* of *connectionist* cognitive science. The computational operation of PDP systems, in contrast to conventional systems, is inextricably bound-up with a theory of content determination. Specifically, the manner in which PDP systems execute computational operations means that connectionism is committed to a structural isomorphism TCD. What is required to prosecute this claim is, of course, a detailed investigation into the classical and connectionist frameworks. Such an examination begins in subsection 1.1., and is pursued, at a deeper level, in subsections 1.2. and 1.3. The argument commences, however, not with an account of how classicism and connectionism differ, but with consideration of what they have in common.

1.1. Generic Computationalism

What does it mean to say that classicism and connectionism are both *computational* theories of cognition? Why are conventional computational systems and PDP systems both computational devices? What is required is an account of computation which is neutral between classicism and connectionism; an account, that is, of *generic* computationalism.[116] This is the issue I will address in this subsection. My strategy will not be to attempt an exhaustive survey of existing characterisations of computation. Rather, I aim to settle on a philosophical account which will be specific enough to distinguish computational devices from other complex systems, yet broad enough to satisfy philosophers of both the classical and connectionist persuasion. With the characterisation of generic computationalism in place, I will be in a position to precisely locate the difference between conventional and PDP accounts of how computational processes are physically realised, and say how this difference, in turn, relates to the issue of content determination.

---

[116] The idea of generic computationalism is in line with the recent recognition that computation is a broad notion, one which subsumes both conventional and PDP devices. See, for example, Clark (1998), Cummins and Schwarz (1991), Dietrich (1989), O'Brien (1998), and Von Eckardt (1993).

A recent attempt to characterise the notion of a computer in such a way as to incorporate both conventional and PDP devices is to be found in Von Eckardt (1993). She writes:

> A computer is a device capable of automatically inputting, storing, manipulating, and outputting information in virtue of inputting, storing, manipulating, and outputting representations of that information. These information processes occur in accordance with a finite set of rules that are effective and that are, in some sense, in the machine itself. (1993, 114)

This passage will be used as my starting point. I will show how both conventional and PDP systems satisfy this characterisation, and why it is neutral between classicism and connectionism. With a little unpacking of the passage, I will then arrive at a formulation of generic computationalism.

First, and most obviously, a computer is a device which employs *representation bearers.* A computer has internal items (objects or states) which represent other entities, internal items which are the bearers of the information the device inputs, stores, manipulates, and outputs. Computation, that is, requires a computational medium: a system of representation bearers (Fodor 1975, 27). Both conventional and PDP systems employ representation bearers. Classicism and connectionism are both representationalist theories of mind. Generic computationalism about the mind is then, for one thing, the view that the mind harbours a vast arrangement of neurally-realised representation bearers. Clearly, this characterisation is, as intended, neutral between classicism and connectionism since it does not specify the nature of the representation bearers realised in the brain.

Second, a computer is a device which automatically inputs, stores, manipulates, and outputs information by inputting, storing, manipulating, and outputting representation bearers. The representation bearers physically realise the information; informational operations are performed via operations on representation bearers. It is not difficult to see how this part of the general characterisation of a computer is met by conventional computers. Conventional computers have separate components for inputting, storing, manipulating, and outputting the information realised by representation bearers (symbols or data structures).[117] The informational operations of these components are automatically performed by means of the machine's basic operations. Such basic operations are defined over the representation bearers.

PDP devices also clearly input, output, and manipulate information by operations "on" representation bearers (patterns of activation, and connectivity matrix dispositions). Informational operations are effected by the creation and transformation of patterns of activation across sets of processing units. A standard PDP network consists of a layer of input units, a layer of hidden units, and a layer of output units. Upon presentation of a cuing input, a pattern of activation is generated across the input layer. Activation spreads throughout the network: activity across the input units is transformed into a certain pattern of activation across the hidden units, which itself gives rise to a pattern of activation at the output layer. The output pattern constitutes the network's "solution" to the "problem" posed by the cuing input. PDP systems also store information. However, as is often remarked, they do so in a fashion very different to that of conventional computers. Rather than having stored information located in a separate component, a PDP system stores information in a distributed fashion, in its pattern of connection weights (see, for example, McClelland, Rumelhart, and Hinton 1996, 32; Clark 1993a, 39; Von Eckardt 1993, 137). In PDP systems there is no sharp division between memory and processing, since it is the network's configuration of weights which influences the course of processing.

So, both conventional and PDP devices satisfy that part of the general characterisation of a computer with which we are currently concerned. And again, as intended, the characterisation is neutral between classicism and connectionism since it does not specify how the informational operations on representation bearers are carried out. I will use "processes" to cover talk of "inputting, storing, manipulating, and ouputting" representation bearers. A computer, then, is a device which *processes* representation bearers. I can now add another element to the formulation of generic computationalism. According to generic computationalism, cognitive processes consist in the *processing of representation bearers*.

I turn now to what I isolate as the third element in Von Eckardt's general characterisation of a computer. This is the claim that the informational operations

---

[117] *Classicism* is not necessarily committed to a discrete *location* for memory. Nonetheless, the processes subserving memory are distinct from those subserving other informational operations.

subserved by the processing of representation bearers "occur in accordance with a finite set of rules that are effective and that are, in some sense, in the machine itself" (Von Eckardt 1993, 114). Again, this part of the characterisation is clearly satisfied by conventional computers. The rules which govern the operation of conventional computers are, as everyone knows, computer programs. Such rules are effective because, in Von Eckardt's words, "they are built out of the basic instruction set of the virtual machine, and this basic instruction set is, in turn, ultimately reducible to the effective basic instruction set of the computer's machine language" (1993, 125). A conventional computer is either a stored-program machine or a wired-program machine. In both types of machine, of course, the program is "in the machine itself". In the former case, the program is stored in the machine's memory (as lists of symbolically encoded instructions); in the latter, the program is wired into the machine.

Indeed, so clearly do digital computers conform to the requirement that their operation be in accordance with internal rules, that classicism is often *characterised* as the view that cognitive processing is *rule-governed.* One would be hard pressed to do better than "rule-governed symbol manipulation" as the most succinct characterisation of classicism. Such talk of "rule-governed" operation is firmly entrenched and entirely befitting; it captures something distinctive about classical cognitive science (in that the operation of a PDP system is not governed by a distinct set of computational rules). This creates a hitch in the attempt to arrive at the most appropriate generic characterisation of a computational device and computationalism. Because of its close association with a fundamental feature of classicism, talk of "rules" does not seem to express the generality, or neutrality, sought. This hitch can be easily overcome though, with a minor modification to Von Eckardt's terminology. I propose to speak of the processing of representation bearers being "*disciplined*", rather than processing occurring in accordance with a set of "rules". Such terminology avoids any temptation to think only of the way in which informational operations are effected in classical cognitive science.

So, what is it that disciplines the working of a PDP system such that it is able to effect informational operations? This is what Von Eckardt has to say:

> Once [a PDP] network has been trained, it produces the output representation automatically, given the input (and., of course, a pattern of connectivity among the units) in accordance with a set of effective,

> internalised rules (namely, the *transition and output functions governing the operation of each individual unit*). In contrast to the high-level rules in a conventional machine, these rules are not defined "over" the high-level input representations of the connectionist network. Nevertheless, the process of producing an output when given an input is completely rule-governed in the sense that C2 [that is, Von Eckardt's general characterisation of a computer, quoted above] requires. (1993, 137; emphasis added)

So, according to Von Eckardt, the "mechanism" which disciplines a PDP system's operation, such that it performs information processing, consists in its transition and output functions. These functions apply to the operation of the network's individual processing units. The transition function determines the unit's response (that is, its new level of activation) to input it receives. The transition function takes the activation value of the unit and the net input to the unit at t, and gives a new activation value for $t + 1$. The output function takes the unit's activation value and gives a value for its output signal (Rumelhart, Hinton, and McClelland 1986, 48-52; Von Eckardt 1993, 129).

In my view, however, Von Eckardt is not correct in claiming that it is the transition and output functions which constitute the disciplinary mechanism of PDP systems. I am not suggesting that these functions are unimportant to an understanding of the operation of PDP networks. In order to understand a network's behaviour, we have to understand the operation of its constituent processing units, since the activity of each unit contributes to the behaviour of the network as a whole. But, in understanding the behaviour of a network *qua information processor*, to focus on the transition and output functions is to focus on the wrong level of analysis. A computer, as indicated above, is a device which processes information by processing representation bearers. In a PDP system, both representation bearers, and the processing of representation bearers, are network-level features. A PDP system processes information by the successive transformation of activation patterns over its input units into activation patterns over its output units. Since the disciplinary mechanism must apply to the processing of representation bearers, then because such processing is a network-level feature in a PDP system, the disciplinary mechanism must also be understood at the network level.

The network-level feature to which processing accords in a PDD system is its *pattern of weighted connectivity*. The course of a processing episode in response to a particular input is governed by the network's particular configuration of weighted connection lines. It is this pattern which dictates the particular output pattern produced in response to a particular input. Now, *precisely* how a network's pattern of weighted connectivity governs its operation as an information processor is something I'll examine in more detail below (subsec. 1.3.).

Both conventional and PDP systems, then, satisfy the third element of Von Eckardt's general characterisation of a computer (though, in regard to the latter, not quite in the way Von Eckardt thinks). The operation of each type of device is disciplined so as to effect information processing. Again in keeping with the aim for a characterisation neutral between classicism and connectionism, the nature of the disciplinary mechanism is not specified. So, a computer is a device whose operation consists in the *disciplined* processing of representation bearers. I can now add a third element to the characterisation of generic computationalism. According to generic computationalism, cognitive processes consist in the *disciplined processing of representation bearers.*

So, Von Eckardt's general characterisation of a computer is both satisfied by, and neutral between, conventional and PDP devices. From this characterisation I have derived a formulation of generic computationalism. But does this characterisation of a computer, and the associated formulation of generic computationalism, capture what is at the heart of computation? The answer, in short, is not quite. The formulations, I think, require a little more precision. As I've spelt-out, computation involves the processing of information by the disciplined processing of representation bearers. What does such an operation amount to? What is it about the processing of information by the disciplined processing of representation bearers that makes for *computation*?

A computational operation is one where the disciplinary mechanism governing the processing of representation bearers ensures that representation bearers are processed in a *systematic* fashion. The representation bearers which participate in a computational process exhibit non-arbitrary semantic relations to one another. That is, the manipulation of representation bearers is disciplined so as to engage representation bearers which bear comprehensible semantic relations to one another. All this can be summed up by saying that computation is an operation whereby the causal relations among representation bearers are *"semantically coherent"*.

The term "semantic coherence" is due to Fodor (1987, 20). Like so many of the things to which Fodor draws our attention, the notion of semantic coherence has found its way into the wider cognitive science community, and its importance is being increasingly recognised. O'Brien (1998) liberalises Turing's conception of computation to arrive at a generic conception, a central element of which is the idea of semantic coherence. Clark (1993a, 4-5) talks of the "semantic good behaviour" of physical systems such as calculators, "automatic symbol crunchers" (that is, digital computers), and cognitive agents. Haugeland (1985), in his exposition of digital computation, writes of "coherence" and of "making reasonable sense". And Horgan and Tienson (1996, Chps. 6 and 9) speak of "content-appropriateness".

The notion of semantic coherence is difficult to precisely define, though it is possible to get a workable intuitive feel for the idea. The example Fodor uses is truth-preservation. In the context of espousing the digital account of computational operations, he writes:

> ...the machine [that is, the digital computational device] is so devised that it will transform one symbol into another if and only if the propositions expressed by the symbols that are so transformed stand in certain semantic relations – e.g., the relations that the premises bear to the conclusion in a valid argument (1987, 19).

Similarly, in discussing Turing's insight into how to mechanise computational operations, Fodor says:

> "I'll bet", Turing (more or less) said, "that one could build a *symbol manipulating machine* whose changes of state are driven by the material properties of the symbols on which they operate (for example, by their weight, or their shape, or their electrical conductivity). And I'll bet one could so arrange things that these state changes are *rational* in the sense that, given a true symbol to play with, the machine will reliably convert it into other symbols that are also true (1992, 6, latter emphasis added).

Clark provides the example of arithmetic calculation:

> It is the lot of some parts of the physical world to be semantically well behaved. A pocket calculator (to give a mundane example) is semantically well behaved with respect to the arithmetic domain; its

> physical state transformations, when subjected to a specific and static kind of interpretation, turn out to track semantically sound (i.e. correct) arithmetical derivations (1993a, 4-5).

Significantly, as shown in Chapter One (subsec. 1.3.), folk-psychological explanations of behaviour attribute semantically coherent sequences of mental states, such as: if a cognitive agent desires state of affairs *S*, and believes that behaviour *B* will facilitate *S*, then (*ceteris paribus*) the agent will form an intention *I* to perform *B*. In general, then, the notion of semantic coherence is that of causal interactions between semantically interpretable items (representation bearers) which respect or preserve semantic properties.

I arrive, then, at the final general characterisation of a computer, and the final formulation of generic computationalism.

> *A computational system is a device whose causal operation is disciplined such that it processes representation bearers in a semantically coherent manner.*

Embedded within this account is, of course, a characterisation of generic computation.

> *Computation is a disciplined operation whereby representation bearers are processed in a semantically coherent manner.*[118]

This then provides a characterisation of generic computationalism about the mind.

> *Generic computationalism about the mind is the view that cognitive processes are disciplined operations whereby neurally-realised mental representation bearers are processed in a semantically coherent manner.*

These characterisations will figure centrally in the discussion of the conventional and PDP accounts of the physical realisation of computational processes which follows shortly.

The recognition that at the heart of computation lies the idea of semantically coherent processing provides a useful perspective on the appeal of understanding cognitive processes as computational processes. It is, as shown in Chapter One, a "deep fact" about the cognitive mind that there obtains a "parallelism between content and causal relations" (Fodor 1987, 14, 19). That is, there exists a correspondence between the content relations among mental states and the causal relations among mental states. This is the *rationality* of the cognitive

---

[118] For an almost identical definition of (generic) computation see O'Brien (1998).

mind. Understanding cognitive processes as computational processes is attractive because a computation is an operation where there exists a parallelism between causal relations and content relations among representation bearers. Such a parallelism is the notion of semantically coherent processing. A device's causal operation is semantically coherent insofar as there is a *functional isomorphism* between representation bearers and their representational objects. The causal network of representation bearers is isomorphic to, or mirrors, certain "physical-value" relations among their representational objects. The appeal of cognitive science lies with the fact that such a functional isomorphism can hold between the causal operation of a computational device and certain real relations which obtain among the external objects of its cogitations. And this suggests a naturalistic solution to the rationality problem. That is, since a computational device *mechanistically* processes representation bearers in a semantically coherent fashion, viewing the cognitive mind as engaging in computational operations offers an answer to the question of how rationality is mechanically possible. Indeed, when it comes to the rationality problem, computationalism is currently the only game in town.

1.2. Classicism: Computation and Content Determination

In this subsection, I will discuss the conventional (that is, digital) means of physically realising computational operations, and the classical account of the computational nature of cognitive processes.[119] Of particular importance for the line of argument being pursued over this section is the manner in which the processing of representation bearers is disciplined in conventional computational systems. The discussion of conventional computation will highlight why a TCD is not part of the classical means of generating semantically coherent operations. The implication is that classicists are not, for reasons to do with the task of securing semantically coherent processing, committed to a particular TCD. This result sets-up the contrast with the main conclusion to be reached in subsection 1.3., namely, the claim that

[119] My discussion of classicism draws upon a number of sources: Block (1990), Clark (1989), Copeland (1993), Dennett (1987), Fodor (1975, 1980, 1987), Fodor and Pylyshyn (1988), Haugeland (1985), Horgan and Tienson (1996), O'Brien (1993, 1998), and Pylyshyn (1984). Direct reliance on a particular author is acknowledged in the usual fashion.

connectionists are committed to a TCD as the mechanism for the disciplined processing of representation bearers.

1.2.1. Conventional Computation and Classicism I: Representation and Processing

According to the formulation of generic computationalism arrived at above, computation is a disciplined operation whereby representation bearers are processed in a semantically coherent fashion. The first step in arriving at this formulation consists in the commitment to representationalism. That is, computation presupposes a representational medium, an arrangement of representation bearers. The first step in showing how to physically realise computational operations, then, is an account of how to physically realise information; an account, in other words, of how to physically realise a representational medium.

In general terms, constructing or realising a representational medium consists in providing a semantic interpretation of some variable physical property. In digital or conventional computation, a *symbolic* representational medium is employed. A symbolic representational medium is physically realised by first systematically partitioning some continuously variable physical property so as to create a set of simple token types. Such partitioning generates *syntactic* structure. The partitions are then provided with a semantic interpretation. This involves specifying a systematic mapping from the syntactically defined simple token types to objects that these token types are taken to be about, or represent. The result of such an interpretation is the specification of a set of simple (or atomic) symbolic representation bearer types. The systematically partitioned physical property becomes, under interpretation, a symbolic representational medium. Each symbol type has a particular syntactic structure and a particular semantic value, or content. In addition, the ways in which the simple symbol types can be combined to form complex (or molecular) symbol types is specified. Each complex symbol type has a composite syntactic structure and a composite semantic value. The former is a function of the syntactic structure of its atomic constituents, together with the manner of their combination. The latter is a function of the semantic values of its atomic constituents, again together with its constituent structure. The complex symbol types have, that is, a combinatorial syntax and semantics. According to classicists, the digital way of realising a representational medium is the brain's way. The classical computational theory of mind holds that the brain realises a system of symbolic mental representation bearers. Classicism, that is, posits a neurally-realised *language of thought* (see Fodor 1975, 1987, and Fodor and Pylyshyn 1988).

The second step in arriving at the formulation of generic computationalism concerned the processing of representation bearers. The second step in showing how to physically realise computational operations, then, is an account of how to physically realise the processing, or manipulation, of representation bearers. The story here, of course, overlaps with the first step above. Information is physically realised by representation bearers, and representation bearers, as physical objects, can be physically manipulated. Moreover, physical differences between representation bearers mean that different types of representation bearers can be processed in different ways. Defining representation bearer types, and thereby their differential causal roles, requires a specification of what physical variations are to constitute a difference in representation bearer type. In digital computation, not all physical variation in the relevant physical property of the representational substrate constitutes a difference in representation bearer type, and hence a difference in the causal role of a representation bearer. As stated above, the continuously variable physical property of a digital system's representational substrate admits of a systematic partitioning. It is such partitions (and their concatenations) which, under interpretation, constitute representation bearer types. This means that the mechanisms which detect and manipulate representation bearers do so not in terms of the representation bearers' microphysical details, but rather according to a higher-level physical property they possess. This higher-level property is, of course, *syntax*. Different types of representation bearers have different causal roles by virtue of their distinctive syntactic properties; there obtains a parallelism between the syntactic properties of symbolic representation bearers and their causal roles. In other words, it is to such syntactic properties that operations on representation bearers apply.

The first two steps of the conventional realisation of computational processes mean that a digital device is a *syntactic engine*, a physical device whose causal operation is governed by the syntactic features of the symbolic representation bearers resident in its representational medium. According to classicists, our cognitive processes are the operations of neurally-realised syntactic engines.

1.2.2. Conventional Computation and Classicism II: Semantic Coherence and Content Determination

The appeal of a syntactic engine lies with the fact that it goes some way in showing us how a purely physical device could behave *as if* it were a semantic engine. A physical device behaves as if it were a semantic engine when its causal operation respects the semantic properties of its physically realised representation bearers. But, by itself, a syntactic engine does not behave as if it were a semantic engine. There is nothing in the syntactic features of symbolic representation bearers themselves which dictates that they will be processed in a manner which guarantees sensible semantic relations among them. To see this, notice that conventional systems require the employment of composition rules which govern the concatenation of simple symbols to yield complex symbols. Such composition rules permit only those concatenations which result in a semantically reasonable composite semantic value, one, that is, which makes sense. A syntactic engine, then, is not yet a computational device.

It is at this point that we must return to the third step in the formulation of generic computationalism, namely, the idea that a disciplinary mechanism governs the processing of representation bearers. I claimed that what it means to say that a computation consists in the processing of information is that representation bearers are processed in a semantically coherent fashion. It is the disciplinary mechanism which secures the semantically coherent processing of representation bearers.

In digital computation, the mechanism which governs the processing of representation bearers consists in a set of state transition *rules*. These rules govern processing by being applied to the syntactic properties of symbolic representation bearers. And it is these rules which ensure that symbols are processed in a semantically coherent manner. That is, though they are syntactically applied, the rules are *there* in order to secure semantic coherence. Only a *rule governed* syntactic engine is a digital *computational* device. In having its causal operation governed by rules – rules which are framed so that operations on symbols respect their semantic properties – a syntactic engine behaves as if it were a semantic engine.

These rules, of course, must themselves be physically realised. In a stored-program digital computer, state transition rules (that is, programs) are explicitly encoded by symbols in the machine's representational medium. This is not to say that digital computation *per se*, and thereby classicism, is committed to the view that state transition rules *must* be explicitly represented (see Fodor 1987, 25). Rather, classicists find the explicit encodement of rules attractive because it yields devices which are programmable. That is, the way the device operates can be altered by altering the explicitly encoded state transition rules. This confers significant flexibility of operation; stored-program machines can perform a range of different computational tasks. By explicitly encoding different sets of state transition rules, a stored-program machine can be transformed into any one of a large number of possible special-purpose, or virtual, machines. The fact that rules can be explicitly encoded, then, means classicists can appeal to this idea of a virtual machine in order to explain how we learn to perform a particular cognitive task. Classicists are able to posit that learning a natural language, for example, is a matter of explicitly encoding a set of state transition rules which transform part of the brain into a language-processing virtual machine.

At bottom, however, even a stored-program device must have certain primitive rules built into its basic physical structure. That is, components of the system have a physical organisation which embodies a set of primitive computational instructions; the components behave as if in accordance with a set of state transition rules. In contrast to the rules explicitly represented in the device's representational medium, the information in these primitive instructions is "tacitly" represented (Dennett 1987, 218). In standard digital machines, primitive instructions are embodied in groups of transistors, wired together to comprise read/write "gates". [120]

According to classicism, again, the digital way of realising computational operations is the brain's way. Certain rules which govern the processing of symbols are physically realised in the brain's representational medium. Other rules

---

[120] I am indebted to O'Brien (1993, 1998) for the discussion of the physical realisation of rules in conventional computational systems.

implicated in processing – the primitive computational instructions – have, through evolutionary forces, been built into the structure of the brain's basic mechanisms.

I reach now a point of consolidation. I've traced a route from generic computationalism to the digital realisation of computational processes, and thereby to the classical computational theory of mind. I've examined the particular story digital computation tells about the physical realisation of representation bearers, the processing of representation bearers, and the disciplinary mechanism which secures the semantically coherent processing of representation bearers. The result of this story, when applied to the mind, is classicism. According to classicism, cognitive operations consist in *syntactically*-applied *rules* governing the processing of mental *symbols* in a semantically coherent manner. Or, a little more elegantly, cognitive processes consist in the rule-governed, syntactic manipulation of mental symbols.

Nothing I've said about the classical doctrine departs from the orthodox understanding: classicism is typically characterised as rule-governed symbol manipulation. Where the above discussion of both computation and classicism differs from the usual treatments is the focus placed on semantic coherence and the way it is secured. I have stressed that at the centre of the notion of computation lies the idea of semantic coherence. And it is the manner in which semantic coherence is achieved which is fundamental to a computational theory. On Horgan and Tienson's (1996, 24-29) view, for example, the commitment to rules – "programmable, representation-level [PRL] rules", as they call them – is, in a sense, the most fundamental assumption of classicism. Describing such an assumption, they write: "Cognitive processing conforms to precise, exceptionless rules, statable over the representations themselves and articulable in the format of a computer program" (1996, 24). They go on to say: "...whenever we talk about classicism's contention that cognition is subserved by computation, we will mean computation over representations, i.e., *processing that conforms to PRL rules*" (1996, 25, emphasis altered). Horgan and Tienson are right to emphasise the importance of rules to classical cognitive science. The sense in which the commitment to rules is the most fundamental assumption of classicism is that it is this assumption which goes to the notion of semantic coherence. If, as I have suggested, semantic coherence lies at the heart of computation, then that part of the classical doctrine which concerns semantic coherence is the most fundamental feature of classicism. In conventional computational systems, as suggested above, it's the rules which secure semantically coherent processing. For classical cognitive science, that is, the source – the *only* source – of semantically coherent behaviour, is the rules which discipline the processing of symbolic representation bearers.

Where the discussion of classicism also departs from the usual treatments is the interest with the relationship between disciplined processing and the question of content determination. The upshot of the above discussion is that the question of content determination is not an issue classicists have to broach in their account of cognitive processes, in, that is, their account of semantically coherent processing. The question of where the content of mental symbols "comes from" – the question of *in what manner* mental symbols have their content conferred – is beside the point, since it is the rules which ensure that symbols are processed in a manner which respects their content. And so I reach the main conclusion of this subsection. Classicists are not forced to defend any one TCD over another. Classicists, at least insofar as the task of securing semantically coherent processing is concerned, are free to adopt any one of the currently available TsCD.

1.3. Connectionism: Computation and Content Determination

This subsection does for PDP, and thereby connectionism, what the preceding subsection did for conventional computation and classicism.[121] As with the discussion of classicism, I will break the examination into the PDP account of computational operations into two further parts. Employing the formulation of generic computationalism, I'll examine the particular story PDP tells about the physical realisation of computational operations, and thereby the connectionist account of the computational nature of our cognitive processes. Of central importance will be my claim concerning the disciplinary mechanism which underwrites the semantically coherent behaviour of PDP systems. I will demonstrate why the issue of content determination, and, in particular, the structural isomorphism TCD, is implicated in the connectionist account of cognitive processes. Connectionism, it is widely assumed, is a computational theory. But despite the enormous amount that has been written on connectionism – in

[121] My discussion of connectionism is informed by a number of sources: Bechtel and Abrahamsen (1991), Clark (1989), Churchland (1990), Horgan and Tienson (1989, 1996), McClelland and Rumelhart (1986), O'Brien (1998), and Von Eckardt (1993). As with the discussion of classicism, more direct dependence is acknowledged in the usual fashion.

particular, on how it differs from classicism – few have examined the question of just what it means to say that connectionism a species of computationalism.[122] This unsatisfactory situation is something I aim to rectify. I am going to argue that structural isomorphism does for the PDP computational framework what the state transition rules do for conventional computation, namely, ensure that representation bearers are processed in a semantically coherent manner. Connectionists need to appeal to structural isomorphism – structural isomorphism is a TCD "for" connectionism – because it is structural isomorphism which renders connectionism a *computational* theory.

A strong suggestion that something along the lines of a structural isomorphism TCD fits naturally with connectionism is to be found in the fact that PDP systems are often described as exhibiting a "*semantic metric*" (for example, Clark 1993a, 19. See also Horgan and Tienson 1996, 61-62, 156, though they don't employ the term "semantic metric"). The idea of a semantic metric points toward the structural isomorphism TCD, though this is not generally recognised by those who highlight the notion. Clark puts the idea as follows:

> The semantic (broadly understood) similarity between representational *contents* is echoed as a similarity between representational *vehicles* [representation bearers]. Within such a scheme, the representation of individual items is *nonarbitrary*. ...[An item is] represented by a vehicle...which reflect[s] its position in the relevant similarity space (1993a, 19).

Similarly, Van Gelder writes:

> ...the particular [activation] pattern used to represent an item is determined by the nature of that item, and so similarities and differences among the items to be represented will be directly reflected in similarities and differences among the representations [representation bearers] themselves (1991, 41).

Despite it being often remarked that the generation of a semantic metric is a feature of (trained-up) PDP systems, I know of only one discussion where this feature is

---

[122] Those who have considered the issue include Cummins and Schwarz (1991) and O'Brien (1998).

pressed into the service of a TCD.[123] As I commented above, an enormous amount has been written on the impact of connectionism, yet the question of content determination has remained curiously untouched by this new computational framework. Under the sway of the classically-generated assumption that content determination is orthogonal to cognitive science, connectionists who have stared the idea of a semantic metric in the face have generally failed to appreciate its true significance, interpreting it merely in terms of the *realisation* of content. This is a situation I seek to change. Below I aim to put some flesh on the bare bones of the idea expressed in talk of a semantic metric. I am going to argue that the notion of a semantic metric be understood in terms of the *determination* of content, rather than just, as is usual, in terms of the realisation of content.

1.3.1. PDP-Style Computation and Connectionism I: Representation and Processing

According to generic computationalism, one essential dimension of a computational theory is the representationalist assumption. The PDP approach, as recognised by friends and foes alike, certainly involves such a commitment. Like conventional computation, there is a systematic mapping between the representation bearers of PDP systems and the objects of the represented domain. In contrast to conventional computation, however, the representation bearers of PDP systems are not symbolic entities. Rather, as is widely recognised, PDP systems employ two kinds of "non-symbolic" representation bearers: *activation patterns*, and *connectivity matrices*.

The more common (and more plausible) class of PDP system employs a *distributed* style of representation. That is, the activity of individual processing units (in contrast to those PDP systems whereby information is encoded in a "localist" fashion) does not admit of a discrete semantic interpretation. Rather, it is the collective activity of a subset of units that is semantically evaluable. Such

[123] The discussion is due to O'Brien (1998). Indeed, the line of argument conducted over subsections 1.3.1. and 1.3.2. below has benefitted from O'Brien's paper. In an earlier draft of this chapter, I argued that the structural isomorphism TCD had a natural affinity with connectionism (because of the semantic metric exhibited by PDP systems), and, relatedly, that structural isomorphism was attractive because it showed how transactions among representation bearers could be disciplined in PDP systems. So structural isomorphism was an appealing option for the connectionist. What O'Brien's discussion impressed on me was that structural isomorphism is the correct account of how semantically coherent processing is effected in PDP systems. The structural isomorphism TCD, in other words, is not just an appealing option for connectionists, it is mandatory.

patterns of activation constitute one kind of representation bearer in PDP systems. An activation pattern is simply the simultaneous individual activation levels of a subset of processing units.

Activation patterns are sometimes referred to as *activation vectors*. This is because any activation pattern can be described by a n-tuple of numbers, where n is the number of processing units and each number gives the activation value of a processing unit. Activation patterns can also be characterised as points in a n-dimensional *activation space*. Each dimension, or axis, of the space represents the variable activation level of a processing unit in the relevant subset of units. Any single point in the space has a unique set of n coordinate values, and so describes the simultaneous activation levels of each processing unit in the representing subset of units. This is a perspicuous method of illustration, particularly when it comes to the similarity and difference relations among a system of activation patterns. Activation patterns are (nonsemantically) type-identified in terms of their physical profile, or "shape"; that is, the particular configuration of activity across a set of processing units. It is this physical profile which locates an activation pattern in activation space. Tokens of the same activation-pattern type are depicted by the same point in activation space. The similarity and difference relations activation patterns bear to one another is depicted by relative position in activation space, such that the greater the degree of similarity the closer the location in activation space.

Because each processing unit of a network has a range of possible activation values, the one set of processing units can generate a range of activation patterns, and so represent a range of objects. Or, I should say, at any one time the same set of processing units can generate any one of a range of activation patterns. Activation patterns are transient features of PDP systems. Change the input to the network and the activation values of its processing units change, thereby generating a different activation pattern with different representational content. Activation patterns coincide with what Horgan and Tienson call the "occurrent" possession of content. A system possesses content in the occurrent sense when "a token of the relevant…state type occurs as a concrete event or state within the [system]" (1996, 163). Similarly, activation patterns are said to represent their contents in an "explicit" fashion (see, for example, O'Brien and Opie 1999a). An activation pattern is a distinct, physically structured object which admits of a discrete semantic value; each activation pattern has only one representational object.

The long-term encoding of information in a PDP system is effected by its configuration of connection weights, or "connectivity matrix". A network's connectivity matrix consists in both the weight values on the connections between processing units and the pattern of connectivity. And it is from these two features that "connection matrix", or "connection weight" representation bearers as they are typically called, derive. Two properties of a configuration of connection weights are important for this way of physically realising information in PDP systems. First, connection weights are modifiable. Second, it is primarily the configuration of connection weights which dictates the profile, or structure, of activation patterns a network generates in response to particular inputs. These two properties are implicated in the training regime to which networks are subjected. A network with an initial (and arbitrary[124]) configuration of connection weights is repeatedly presented with a certain input over its input units. Via the repeated application of a "learning rule" the network's configuration of connection weights is successively adjusted.[125] That is, the network is taught to produce a target pattern of activation upon presentation of a certain input. This is known as "training" the network (particular examples of which will be considered below). PDP systems "learn" by such a process of modifying the configuration of connection weights. As a result of its training, a PDP system's connection weights become configured in such a way that the system possesses the capacity to produce a certain activation pattern in response to a particular input. That is, a connection weight representation bearer is best thought of as a *disposition* to generate a certain activation pattern upon presentation of a particular input (see Horgan and Tienson 1996, 166; O'Brien and Opie 1999a).

Of course, in practice a network is trained on not one but a range of different inputs, such that it develops the capacity to generate a different activation pattern in response to each different input. All of the connection weight dispositions

---

[124] In regard to natural cognitive systems, connectionists would claim that evolutionary forces have seen to it that the initial configuration of connection weights is not arbitrary or random. Clearly, there are interesting implications here for the nativism issue. See, for example, Clark (1993a, 1993b).

[125] A commonly employed learning rule is *back-propagation*. A network is presented with an input, and activity spreads through the network resulting in a pattern of activation across the output units. This activation pattern is compared with a target output to ascertain the degree of divergence. On the basis of the amount of error, the connection weights are adjusted so as to generate an output closer to the target output. After repeated applications, the network eventually settles on a configuration of connection weights which produces an output that best matches the target output. See Rumelhart, Hinton, and Williams (1986), and Bechtel and Abrahamsen (1991).

are realised by the one configuration of connection weights. Content is stored in what is called a "superpositional" manner: each connection weight of the configuration is implicated in the storage of every disposition. Because of this, any given disposition (and thereby a distinct content) cannot be identified with an isolable part of the configuration. A configuration of connection weights should be understood as a way of realising content where the one dispositional structure is mapped onto a system of contents, such that each content lines-up with one of the activation patterns the configuration is disposed to produce. In contrast to the explicit content possessed by activation patterns, connection weight dispositions are said to possess content in a "*potentially* active" (Horgan 1997, 12) or *"potentially explicit"* (O'Brien and Opie 1999a) fashion.

While I'm going to argue below that connectionism implicates structural isomorphism on the basis of the computational operation of PDP systems, even at this stage of the discussion certain preliminary considerations to do with the representation bearers of PDP networks suggest that the structural isomorphism version of the resemblance theory is at least compatible with connectionism. The kind of representation bearers deployed by conventional computational devices, on the other hand, appears to be another contributing factor in the contemporary conviction that content determination is orthogonal to cognitive architecture. The *classical* conception of representation bearers as symbols is *neutral* in regard to the orthodox TsCD. That is, understanding representation bearers as symbolic items doesn't seem to demand, debar, or even suggest a particular answer to the problem of content determination (apart from, of course, ruling-out the possibility of a first-order resemblance theory; but this it shares with connectionism). Classicism has dominated cognitive science for so long (indeed, for much of its history, cognitive science *was* classicism), that the possibilities afforded by the connectionist framework are only just now beginning to surface. Cummins (1996) has recently argued for a structural isomorphism TCD over the orthodox approach to content determination (or "meaning as use", as he terms it). In contrast to "use theories", structural isomorphism, for Cummins, has it that meaning is "intrinsic". And as he writes, "what blinded me for a long time to the possibility of intrinsic meaning was the general obsession with symbols" (1996, 93). Symbols don't stand for their representational objects by virtue of their intrinsic properties.[126]

The structural isomorphism TCD requires a representing system which admits of systematic physical variation. More specifically, there must exist systematic physical relations among variations in the physical property which constitutes the *representational medium* of the representational system. The representational medium of a PDP system consists of the patterns of activity across the system's processing units, along with the dispositions to generate such activation patterns stored in the network's pattern of weighted connectivity. Activation-pattern representation bearers exhibit physical relations in terms of what can be roughly characterised as their "shape". The shape of an activation pattern is determined by the particular level of variable activity in each processing unit of the network. It is the similarity and difference relations between the shapes of activation patterns which constitutes the physically-structured representational medium of a PDP system. Importantly, the physically-structured representational medium is directly fused with a physical feature of the network's physical substrate. The activity of each processing unit is a feature of the network's physical substrate, and, as suggested, the various patterns of activity across the network's processing units constitute its representational medium. This is why the relations activation patterns bear to one another are natural physical relations among variations in a property of the network's physical substrate. And it is for this *preliminary* reason that I think structural isomorphism has a natural affinity with connectionism.

I turn now to the *processing* of representation bearers in PDP networks. At the most basic level, the story here is the same as that for conventional computation. Representation bearers physically realise information, and, as physical objects, representation bearers can be transformed. And again, physical differences between representation bearers mean that different types of representation bearers have different causal roles. This, however, is as far as the similarity in the two computational stories goes.

In conventional devices, representation bearer tokens are (non-semantically) type-identified according to their syntactic properties. The causal role

[126] Cummins (1996), however, does not seek to align the structural isomorphism TCD with connectionism.

of a given representation bearer token, then, is defined in terms of its syntax, rather than its microphysical profile. The causal roles of different types of tokens are differentially determined by their distinctive syntactic properties. That is, symbols are "read" and processed on the basis of their syntactic properties. A conventional system, as suggested above, is a syntactic engine. This is due to the nature of the relationship between the device's representational substrate and its representational medium. As shown in the discussion of conventional computation and the structural isomorphism TCD, a conventional device's representational medium exists, so to speak, "at one remove" from its representational substrate; it "transcends" its representational substrate. This means that certain variations in the physical property which comprises the representational substrate are not representationally significant.

In contrast, and as recently highlighted, a PDP system's representational medium is directly "fused" with its representational substrate. Consequently, any variation in the physical property which comprises the representational substrate *is* representationally significant. Consider activation patterns. The profile of an activation pattern is constituted by the simultaneous individual activation levels of the relevant set of processing units. Any variation in the physical property comprising the system's representational substrate – any variation, that is, in the individual activation level of any of the relevant processing units – results in an activation pattern with a different profile, and a new representational value. Tokens of activation-pattern representation bearers are, then, (non-semantically) type-identified according to their *physical* properties. The causal role of any given token is defined in terms of its physical profile. Tokens of different representation bearer types have their differential causal roles by virtue of their differing physical profiles. Activation patterns are "read" through having effects in other areas of the network. An activation pattern "passes" activity to other sets of processing units in the network, resulting in the generation of further activation patterns. The character of this effect is determined by the particular profile of the originating activation pattern (together with the weights of the connections between the participating processing units). Processing in PDP systems consists in the successive spread and transformation of patterns of activation over the input units into patterns of activation over output units. It is the physical properties of the representation bearers which, ultimately, drive the operation of PDP systems. PDP systems, we might say, are not syntactic engines, but rather *physical* engines.

1.3.2. PDP-Style Computation and Connectionism II: Semantic Coherence and Content Determination

A mere physical engine, however – like a syntactic engine without a set of state transition rules – does not behave as if it were a semantic engine of its own accord. As so far described, the PDP framework does not yet, strictly speaking, constitute a computational theory of mind. It is at this point that the following issue must be broached. What governs the *course* of processing in PDP systems? Or, in the familiar terminology of this section: How is the processing of representation bearers *disciplined* such that a PDP system behaves in a semantically coherent manner? It is vital that connectionists address such an issue. At stake is nothing less than connectionism's viability as a robust, alternative computational theory. Let me first briefly reflect on this point.

Classicists and connectionists, for the past decade or so now, have been engaged in a vigorous debate over the relative merits of the two rival accounts of cognition. Amidst the noise of claim and counter-claim the foundational issue of in just what sense connectionism can lay claim to the title of alternative *computational* framework has remained largely unexamined. Philosophers of cognitive science in general, and connectionists in particular, have not always explicitly recognised the central importance of the idea of semantically coherent processing. Once it is out in the open, however, it is clear that connectionism must address the issue.

Classical computationalism has in place a means of securing semantically coherent processing (not the least of its virtues being that it works). As highlighted above, it is a conventional system's set of computational rules that ensures the semantically coherent processing of symbols. The assumption that processing is rule-governed is fundamental to classical cognitive science. Herein, I think, lies the reason why the idea of semantic coherence is rarely isolated and addressed. Classicism builds-in, or incorporates, an answer to the question of how representation bearers are processed in a manner which respects their semantic properties. The commitment to rule-governed processing is so *central* to classical computationalism that the reason why the computational rules are *there* in the first place is typically not highlighted. With the reason rendered explicit, it can be seen that connectionism must provide an alternate account of semantically coherent processing. As is widely recognised, of course, the PDP framework eschews the idea that processing conforms to computational rules. Indeed, for some theorists (most notably perhaps Horgan and Tienson 1989, 1996), connectionism's rejection of rule-governed operation is *the* significant departure from classical computationalism. It therefore behoves connectionists to articulate the disciplinary mechanism which gives PDP systems the capacity to behave in a semantically appropriate fashion.

As far as I can tell, there have been only two attempts to undertake the task just described. The first appears in Horgan and Tienson's (1996) recent elaboration and defence of the connectionist framework.[127] The connectionist account Horgan and Tienson offer rests on what they term *"content-appropriate cognitive forces"*.

> If cognitive processes do not conform to PRL [that is, programmable, representation-level] rules...then what does the causal role of cognitive states consist in? We think connectionism suggests an alternative. Connectionists speak of "spreading activation" in a network. Activation is spread by *forces* within the network. The nodes of an active connectionist representation put out forces that tend to activate (or inhibit) nodes of other representations, and thereby active connectionist *representations* tend to activate (or inhibit) other representations. Furthermore, cognitive states must tend to activate successor states that are appropriate in terms of their content. So the picture suggested by connectionist models is that mental causation is a matter of content-appropriate *cognitive* forces, subserved by spreading activation consisting of subcognitive inter-node forces (1996, 95).

In this passage we see an explicit understanding that cognitive processes, as computational processes, are semantically coherent, or "content-appropriate", operations. The representation bearers constituting cognitive states interact in ways which respect their content. Moreover, Horgan and Tienson clearly think that connectionism's rejection of the classical rule-governed picture of cognitive processes means there is a need for an alternative account of how content-appropriate interactions occur among cognitive states. The idea of "content-appropriate cognitive forces", they claim, provides such an alternative account.

[127] The second is to be found in O'Brien (1998). See note 123. of this chapter.

This idea, suggested by the PDP framework, is, in a way, inevitable. The rejection of the rule-governed picture would appear to leave only one option. PDP systems must, in some sense, *automatically* or naturally engage in semantically appropriate behaviour:

> ...certain kinds of content-relevant interaction are automatic for systems that have states that emit content-relevant cognitive forces. [This is] a key difference from classicist systems, in which the operative PRL rules must determine all such outcomes (Horgan and Tienson 1996, 99).

What Horgan and Tienson's "content-appropriate cognitive forces" picture suggests is that the representation bearers in PDP systems, in effect, discipline themselves.

Of course, as Horgan and Tienson well realise, the idea of content-relevant forces cannot be the whole story. It must be explained how it is that representation bearers in PDP systems come to have such content-appropriate powers. For Horgan and Tienson (1996, 154), the explanation lies with the training of a PDP system, the result of which is a shaping or *moulding of the activation landscape*, and the *strategic, rather than arbitrary placement of representation bearers* on this activation landscape. A training regime moulds a network's activation landscape – renders it with a characteristic "topography" – by bringing about successive modifications to the network's pattern of weighted connectivity. The strategic placement of the representation bearers which constitute cognitive states on an activation landscape means that "their relative-position relations systematically reflect key semantic properties and relations of the relevant individual cognitive states themselves" (1996, 156). The characteristic topography of an activation landscape, and the strategic situating of representation bearers on the activation landscape, "are jointly just right to subserve content-appropriate cognitive transitions...." (1996, 154). So, PDP systems, in contrast to conventional devices, don't have semantic coherence enforced by a set of computational rules. Rather, the training regimes to which they are submitted means that they naturally come to be semantically well-behaved.

Horgan and Tienson's account is, I think, very much on the right track. However, in my view, it doesn't go quite far enough; their proposal requires one further element. In order for Horgan and Tienson's account to work, the PDP framework must be understood as implicating the structural isomorphism TCD. That is, what Horgan and Tienson don't recognise is that the *structural* *isomorphism TCD underlies*, or underpins, their "content-appropriate cognitive forces" proposal. And this means that ultimately it is structural isomorphism which secures the semantically coherent behaviour of PDP systems.

The proposal Horgan and Tienson offer relies on the progressive "tuning" of two processes. During the course of training a PDP system's activation landscape is gradually moulded, and representation bearers are strategically located on this activation landscape. As Horgan puts it in a later work, "the core idea is that the shape of the activation landscape and the positioning of representational points on the landscape are "made for each other" " (1997, 25). What Horgan and Tienson are claiming is that the co-evolution of these two processes results in an alignment of content relations with relative-position relations. But this leaves a crucial question unanswered. Just *how* is such an alignment effected? Why is the strategic placement of representation bearers such that relative-position relations reflect content relations? What is it about the moulding of the activation landscape which means that the judicious placement of representation bearers confers *content-relevant* causal powers?

Consider again the two elements of Horgan and Tienson's proposal. One element is that a PDP network's activation landscape, during the course of training, is successively moulded; it acquires a characteristic topography. The moulding of an activation landscape is the training of a network such that it develops the capacity to generate a certain range of activation patterns, each with a characteristic profile, in response to a range of inputs. The activation landscape's topography is that of (or reflects) the various physical profiles of the activation patterns the system has been trained to produce. The various relations between the points on, or regions of, the landscape constitute the relations activation patterns bear to one another – their similarity and difference relations – in terms of their differing profiles. The other element of Horgan and Tienson's proposal is that activation patterns are located on the activation landscape in a non-arbitrary way, in a manner, that is, whereby relative-position relations coincide with content relations. The claim here is that the co-evolution of these two processes aligns the content properties of representation bearers with their physical properties. But, and here is the shortfall, this does not say much more than what is meant by the idea of representation bearers possessing content-relevant powers. What we are not told is what it is *by virtue of* that the content relations among representation bearers are in-step with their physical relations. How are the various locations on the activation landscape able to systematically embody differential contents? Why is the moulding of the activation landscape such that its topography acquires representational significance?

What is required here, as foreshadowed above, is for Horgan and Tienson's proposal to be augmented with the structural isomorphism TCD. In slightly more forceful terms, what is actually taking place in the moulding of an activation landscape is the emergence of a structural isomorphism between a PDP network's representational substrate and a representational domain. And, at bottom, it is this isomorphism which produces content-appropriate processing. The line of thinking (as regards the claim that Horgan and Tienson's proposal needs augmentation) is this. The actions of representation bearers in a PDP system are not administered by a separate body of computational rules. Rather, the causal powers of these representation bearers depend on their characteristic physical profiles. Now, as described above, the shape of the activation landscape is comprised of representation bearers arranged in terms of their particular profiles. This means that relative positions on the landscape describe relative causal powers. That is, the differing locations of activation patterns on the activation landscape reflect their differential causal powers. So, for representation bearers positioned on the activation landscape to possess *content*-appropriate causal powers, the arrangement of the activation landscape must itself be relevant to, or bound-up with, the determination of content. The activation landscape, in short, must have content-conferring significance. And, as far as I can see, the only way the topography of an activation landscape can be of representational significance is by virtue of being structurally isomorphic to a representational domain. The structural isomorphism TCD is the only theory which maintains that the physical structure within a system of representation bearers is implicated in the determination of representational content.

The above line of argument, as indicated, can be put in a slightly stronger way. It urges not just that Horgan and Tienson's proposal requires supplementation with structural isomorphism, but that it is the structural isomorphism TCD which *ultimately underwrites* the semantically coherent behaviour of PDP networks. That is, a proper understanding of network training must recognise that what is crucial is the development of a structural isomorphism between a network's activation landscape and its representational domain. What occurs during training is that the network's representational substrate undergoes successive modifications, such that the network comes to store, in its connectivity matrix, dispositions to generate certain activation patterns. The characteristic profile of each activation pattern in the response-set comprises the particular topography of a network's activation landscape. For the representation bearers located along this landscape to possess content-appropriate powers, their differing physical properties, and thereby causal powers, must be aligned with their differential contents. The claim here is that such an alignment occurs because the moulding of the activation landscape is *constitutive* of the determination of content. An activation landscape is so shaped during a training regime that there emerges a structural isomorphism between the landscape and a representational domain, and this amounts to the determination of representational content. In other words, during training a PDP network is physically altered such that it comes to embody, in its configuration of connection weights, an *analogue* of its target domain. And when a PDP network's representational substrate has been so configured that it accommodates an analogue of some domain, it acquires the capacity to *represent* that domain. It is, then, the emergence of such an analogue that gives content-conferring significance to the locations on the activation landscape, and thereby confers content-appropriate causal powers to the representation bearers located along it. In the course of training the physical profiles of representation bearers, and their contents, are being determined *together* as a result of physical modifications to the network. Each pattern of activation the network is disposed to generate has content-appropriate causal powers because the determination of both its content and causal powers coincides in its physical profile.

In order to make the discussion more concrete let me turn to some actual PDP models. With a little careful inspection the structural isomorphism TCD can be seen at work in the operation of current PDP systems, and, I claim, it is what's crucial to their successful performance in their respective task domains. That is, the structural isomorphism TCD is actually embodied in the computational operation of extant PDP systems. Structural isomorphism is a TCD "by" connectionism.

I begin with a rather simple PDP system, the mine/rock network developed by Gorman and Sejnowski and discussed by Churchland (1988, 1990). This network is trained to distinguish between the sonar echoes bounced from explosive mines and the sonar echoes bounced from rocks of a similar size to the mines. The task is not straightforward. The difference in the echoes from the two types of object is undetectable to the untrained ear, and even highly trained observers perform less than perfectly. Moreover, every individual echo is different, as all rocks and mines vary in size, shape, and orientation relative to the sonar pulse.

Two sets of echo samples provide the network's learning materials. A large set of differing sonar echoes returned from rocks is collected; likewise, various sonar echoes returned from mines. Each echo in a set is subjected to spectral analysis by being sampled for its relative energy levels at 13 different frequencies. This yields two sets of 13-element vectors. It is these vectors which serve as inputs to the network. The standard feedforward network has 13 input units, a hidden layer of 7 units, and 2 output units (see Figure 6.1.). Input is provided to the network by activating each of the 13 input units to a level corresponding to each of the 13 values comprising the input vector. Activation of the input layer is propagated through the network, transformed at the hidden units, and produces an output across the output units which is described by a 2-element vector. An output vector at or close to (1, 0) indicates that the source of the echo is a mine; an output vector at or close to (0, 1) means that the echo has returned from a rock.

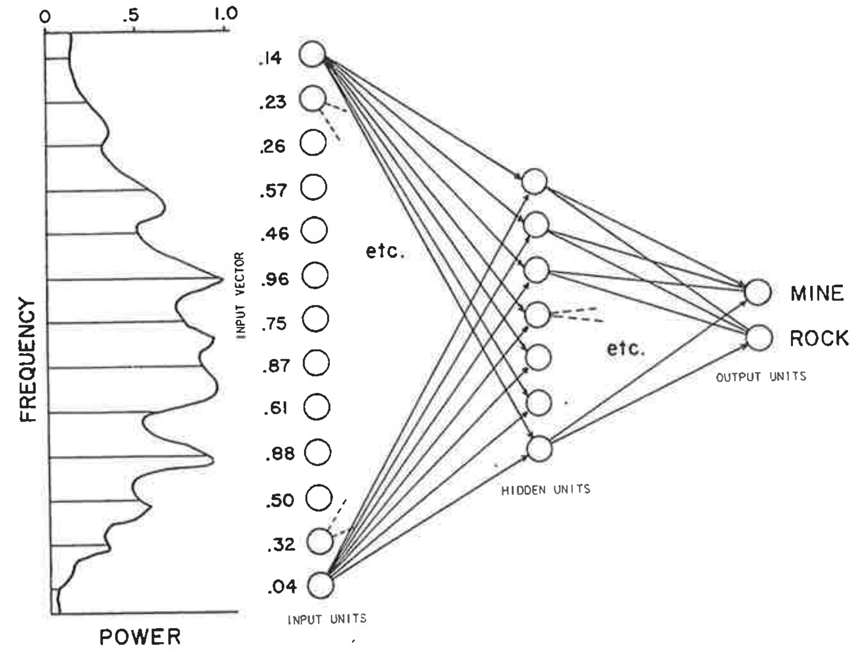

Figure 6.1. A sample sonar echo (input vector) and the architecture of the mine/rock network. (From Churchland 1988, 159).

Network training takes place via the individual presentation of a range of both rock and mine echoes, and the employment of the back propagation learning procedure. Initially, of course, due to the random assignment of weights, the network generates ambiguous output vectors, such as (.49, .51) or (.47, .53). Eventually though, after many hundreds of cycles, the network converges on a configuration of connection weights such that it can, with a high degree of reliability, give an output vector close to (1, 0) when presented with a mine echo, and an output vector close to (0, 1) in response to a rock echo. The network, in short, learns to distinguish mine echoes from rock echoes.

What enables the network to successfully execute this task? One condition which must be satisfied is that the representational domain exhibits some sort of structure. In this case we are, in a sense, dealing with a very simple structure. The representational domain, in some way, admits of a binary division into mine echoes and rock echoes. Some feature, or conjunction of features, which all mine echoes possess differentiates them from rock echoes, and the network, of course, must be sensitive to this difference. On the other hand, that which serves to divide the representational domain into two regions is likely to be (as talk of a "conjunction of features" suggests), some complex of rather subtle and abstract properties. All mine echoes (for example), recall, are superficially different, yet there is no obvious difference between this set of echoes and the set of rock echoes. The difference between the echoes from the two sources is an underlying structural one. What we can call the character or profile of a sonar echo is constituted by variation along each of 13 dimensions (with each dimension admitting of a rather wide range of variation). Mine echoes are similar to one another, and differ from rock echoes, in terms of values along each of these dimensions (or relations between these values). It is the subtle similarity in the profiles of echoes emanating from each type of object which constitutes the structured represented system. And it is the difference between the two sorts of distinctive similarity profiles which is discernible by the network.

What we are really interested in though is how the network is able to detect and respond reliably to this difference. It is the effect the training regime has on the network's organisation of its hidden unit activation space which is crucial to an explanation of its successful performance. As a result of training the network converges on a configuration of connection weights which effects a gross partitioning of its 7-dimensional hidden-unit activation space. The partitioning of activation space is such that the network is disposed to generate, upon presentation of sonar-echo input, an activation pattern which falls into one of two distinct regions of this space. That is, mine-echo input results in an activation pattern across the hidden units with a profile depicted by a point in one region of activation space, whereas rock-echo input produces an activation pattern with a profile depicted by a point in a separate (non-overlapping) region of this space. What is it about these two regions which means that an activation pattern falling within one region results in the output "mine", while activation of the other results in the output "rock"?

The answer is that these two regions of activation space correspond to the development of a *structural isomorphism* between the network's representational substrate and (an aspect of) its task domain. As a result of modifications to the network's pattern of weighted connectivity, it comes to store dispositions to generate activation patterns with profiles of two general kinds. Variation between the two groups of activation patterns (in terms of their distinctive profiles) systematically reflects the variation between the character of sonar echoes returned from the two types of object. Those activation patterns that are clustered together in each of the two regions of activation space stand-in for mines and rocks, respectively, because the network's representational substrate incorporates an analogue of the representational domain (the complex of features shared by mine echoes, and the complex of features common to rock echoes).

Over the course of a training regime, then, successive modifications to the network's configuration of connection weights gradually nudge its representational substrate ever closer to an organisation which mirrors the structure of the representational domain. The trained mine/rock network has an activation landscape which has been so moulded that its particular topography manifests a structural isomorphism to certain abstract features of sonar echoes. The network now stores (in its connectivity matrix) the dispositions to generate a certain range of activation patterns associated with mines, and a certain different range of activation patterns associated with rocks, across its hidden units upon presentation of sonar-echo input. A mine echo across the input units is transformed into a pattern of activation over the hidden units with a certain kind of profile (one which places it in a certain region of activation space, or in a certain area on the activation landscape). A rock echo across the input units is transformed into an activation pattern over the hidden units with a different sort of profile. The network's configuration of connection weights is such that the hidden layer is able to extract from the various activation patterns over the input units that subtle weave of features common to mine echoes, and that subtle weave of features characteristic of rock echoes. Because of the structural isomorphism which holds between the activation landscape and the task domain, activation patterns with one kind of profile stand-in for mines, while activation patterns which exhibit the other sort of profile stand-in for rocks. And this means that representation bearers of each type have content-relevant effects: that type of representation bearer which stands-in for mines generates a "mine" response at the output units; that type of representation bearer which stands-in for rocks produces a "rock" pattern of activity over the output units. The content and causal powers possessed by a representation bearer of a given type coincide in its characteristic physical profile. In sum, then, physical modification of the network – the development of a system of representation bearers with characteristic kinds of profiles – is coextensive with the determination of representational content. Underlying the mine/rock network's discriminative capability is the utilisation of the structural isomorphism TCD. As such, this PDP model is an illustration of my claim that the structural isomorphism TCD is suggested by the PDP framework.

Let me turn now to NETtalk, everyone's favourite example of a PDP model. (The original discussion is to be found in Sejnowski and Rosenberg 1987. See also, for example, Churchland 1990, Churchland and Sejnowski 1989, Clark 1990). NETtalk is a large feedforward network that learns to convert written English text into speech; it transforms graphemic input into phonemic output. The task required of NETtalk is a demanding one, much more so than that confronted by the mine/rock network. While the latter network need only learn to discriminate between echoes from two types of object, NETtalk must learn to distinguish the 79 letter-to-sound correspondences found in the English language. The transformation required is notoriously irregular, with lexical context being an important factor in appropriate phonemic output. The network is not provided with any rules concerning the correct transformation of written words into sounds.

NETtalk's architecture consists of three layers of processing units, with a set of 203 input units, 80 units at the hidden layer, and 26 output units (see Figure 6.2.). The output layer is connected to a speech synthesizer, so as to produce audible sounds. The input units are activated by individual presentations of vector codings for 7-letter segments of text. The desired result is that activation over the input units propagated through the network be transformed at the hidden layer into appropriate phonemic output. NETtalk's initially random configuration of connection weights is successively altered during the training period by repeated applications of the back propagation learning rule. Over the course of its training regime, NETtalk proceeds slowly from nonsense babble to half-recognisable words and finally, after many thousands of cycles, to the production of intelligible speech. The network learns the correct pronunciation of those words presented during its training, as well as those it has not previously encountered. Given the nature of the task domain, this is a considerable achievement.

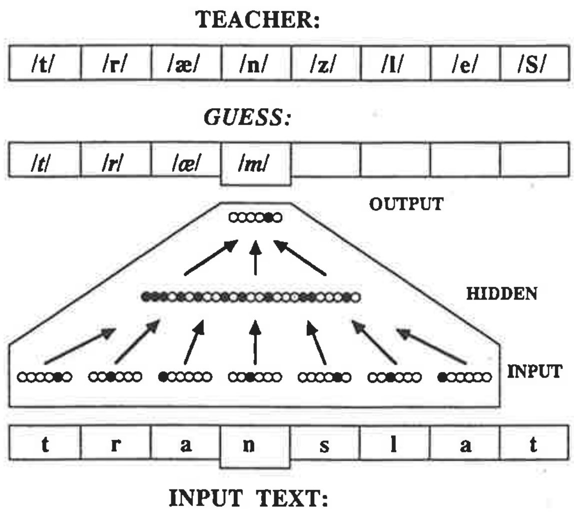

Figure 6.2. Schematic of NETtalk's architecture, showing only a few processing units and connections. (From Churchland and Sejnowski 1992, 118).

As was the case with the mine/rock detecting network, what is crucial to an explanation of NETtalk's performance is the way the network organises its hidden-unit activation space. Analysis of the activity of the hidden units reveals that NETtalk partitions its hidden-unit activation space in such a manner that it admits of 79 distinct regions or clusters, one for each of the 79 letter-to-sound correspondences found in English. Each letter-to-sound correspondence generates an activation pattern with a characteristic type of profile; each such correspondence, that is, activates a distinct region of activation space. As a result of the training regime, then, NETtalk's connectivity matrix is so shaped that it stores a connection weight disposition to produce a certain range of activation patterns for each letter-to-sound correspondence.

Further analysis of hidden-unit activity provides additional insight into how NETtalk executes its task. What is particularly illuminating is the similarity structure of NETtalk's hidden-unit activation space. Sejnowski and Rosenberg (1987) employed a technique known as *cluster analysis* to demonstrate the non-arbitrary nature of the clusters in activation space. Cluster analysis results in a picture of the relationships in activation space between (average) kinds of activation patterns generated in response to certain inputs. What this reveals is a hierarchy of partitions displaying NETtalk's organisation of its hidden-unit activation space so as to accord with the structure of the representational domain (see Figure 6.3.). That is, there exists a systematic relationship between variation along certain aspects of NETtalk's task domain and location in activation space.

The base level of the hierarchy consists of the (small) activation-pattern partitions for the 79 distinct letter-to-sound pairings, and so constitutes the most fine-grained organisation of activation space. And, as suggested, relationships between these partitions are not arbitrary. Activation patterns for letter-to-sound correspondences which are similar fall in neighbouring regions of activation space, while disimilar correspondences produce activation in regions more distantly located. These relationships constitute intermediate levels of the hierarchy, the (mid-sized) regions here lining-up with familiar vowel and consonant subtypes. So, for example, activation patterns for all letter-to-sound pairings employing the letter *a* are located together in a (mid-sized) region of activation space; likewise for pairings employing *e*, *i*, *o*, and *u*, and the relevant instances of *y*. Within the consonants, to provide just a couple of examples, we find those transformations which result in the labial stops */p/* and */b/* to be very closely located, and the same applies for all those transformations which yield the hard-*c* sound */k/*. At the top level of the hierarchy we have the crudest division of activation space. Here the space contains two (large) regions, with letter-to-sound correspondences for vowels clustered in one region, and the letter-to-sound correspondences for consonants falling within the other. At this level NETtalk has partitioned its activation space in accordance with a brute feature of the representational domain, the division between vowels and consonants.

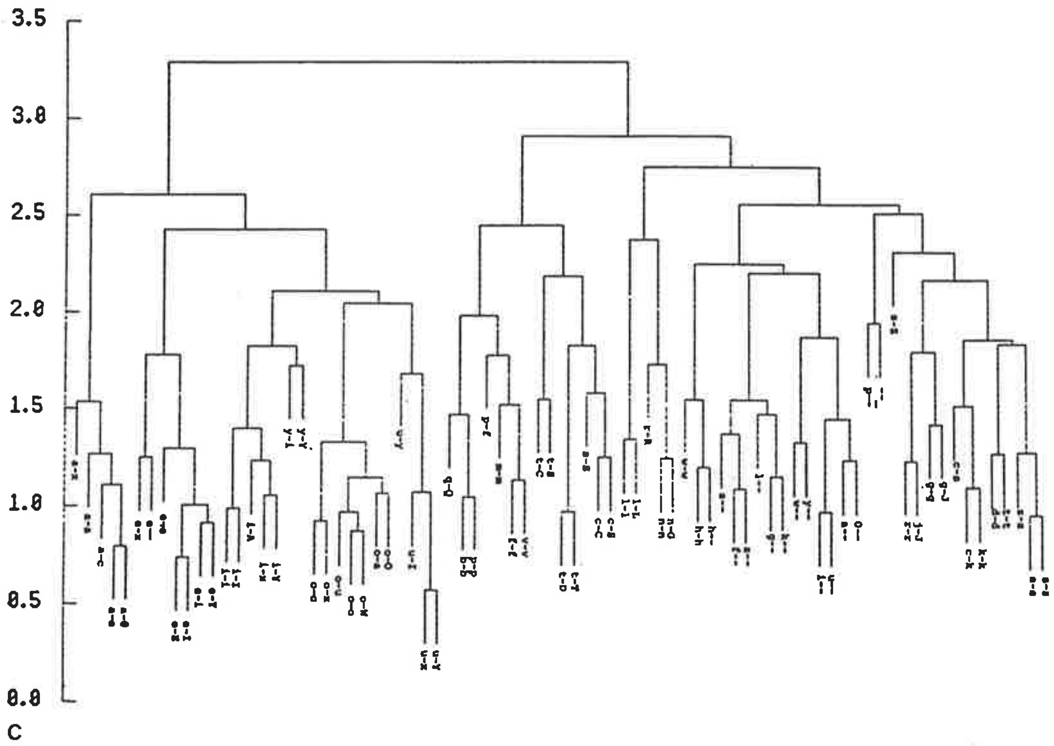

Figure 6.3. Hierarchical cluster analysis of the average activity levels on the hidden units of NETtalk for each letter-to-sound correspondence (l-p for letter "l" and phoneme "p"). The closest branches correspond to the most nearly similar activation vectors of the hidden units. (From Churchland and Sejnowski 1992, 119).

What we discover, then, is that the structure of NETtalk's hidden-unit activation space mirrors the structure of its representational domain. In other words, there is a systematic relationship between variation along certain dimensions of the task domain and variation along a physical property of NETtalk's representational substrate. And so we arrive at a by now familiar place. NETtalk's representational substrate, I propose, incorporates a rather elaborate analogue of its representational domain, the letter-to-sound correspondences found in English. The nature of this analogue is, of course, highly abstract, due primarily to the abstruse character of the representational domain. Unlike the "soundscape" which constitutes the task domain of the mine/rock network, each letter-to-sound correspondence involves a complex blend of features, made up of those elements which together create the sound of an utterance, aspects of the physical shape of written letters, subtle contextual factors to do with the embedding of letters and sounds within syllables and words, and syllable and word boundaries. NETtalk's representational domain is not just a "soundscape", but a highly complex amalgam of "soundscape" and "letterscape". That complexity noted, the claim is that NETtalk's configuration of connection weights, as a consequence of its training regime, has been so moulded that it embodies a particular system of representation bearers. This system of representation bearers exhibits a structural isomorphism to a system of representational objects. The *physical* relations among NETtalk's system of representation bearers correspond with the *physical* relations among letter-to-sound correspondences. NETtalk has the representing system it does, and the constituent representation bearers have the contents they do, because structural isomorphism is the operative principle governing the organisation of its representational substrate. It is by virtue of this structural isomorphism that NETtalk is able to negotiate its task domain. It is, in other words, structural isomorphism which powers NETtalk's computational behaviour.

It is useful at this stage to briefly consider the issue of levels of explanation in connectionist cognitive science.[128] In my view, the level of physical implication (level-3) admits of two kinds of description. At the greatest degree of magnification is the individual activation of each processing unit, and the particular connection weights holding between each processing unit (as well as the transition and output functions being utilised by the network). Taking a more systemic view, there is the collective activity of a group of processing units, along with the overall configuration of connection weights. This perspective is a representational description of level-3, since it reveals the particular representation bearers (the patterns of activation, and the connection weight representation bearers encoded in the pattern of weighted connectivity) implicated in the network's operation. The representational description of level-3 depicts what I'll call the "*intrinsic* physical structure" of the network, the particular, intrinsic nature of the representation bearers it employs. A diagnostic tool such as cluster analysis, in contrast, divulges "*relational* physical structure", the nature of the semantic metric – the analogue – embodied by a PDP network. Cluster analysis provides a level of explanation which is more general than the intricacies of particular, slightly different implementations of the same kind of network. As is well known in PDP circles, PDP systems which are identical in terms of the number and organisation of their processing units, and which are submitted to the same training regime, will – because they begin from a different random assignment of connection weights – end up with *different* activation patterns and connectivity matrices in coming to successfully negotiate their task domain. However, what cluster analysis reveals is that these different implementations end up with the *same partitions* across their hidden-unit activation spaces. My claim, then, is that it is relational physical structure which is common across different incarnations of the same kind of network. The idea is that while the

---

[128] By "levels of explanation" I mean to refer, of course, to the influential tripartite analysis of explanation in cognitive science due to David Marr (1982, 1990). Andy Clark (1990) provides an interesting discussion on the *methodology* of explanation in cognitive science. In barest outline, Clark argues that connectionism effects an inversion of the explanatory methodology deployed in classical cognitive science. Both classicism and connectionism begin, at level-1, with an abstrast analysis of the computational task in question. The classicist seeks to "progressively refine" (1990, 283) a level-1 description until it comprises a level-2 processing story (an explanation detailing the disciplined processing of representation bearers), and so adopts what I call a "top-down" methodology. The connectionist, in contrast, moves straight from a level-1 task analysis to a level-3 implementation (an up and running PDP model). Adopting a "bottom-up" methodology, the connectionist then moves back upwards to a level-2 computational explantion, an account, that is, of the disciplined processing of representation bearers underwriting a network's negotiation of its task domain.

differing incarnations of NETtalk, for example, are likely to represent a certain letter-to-sound correspondence by activation patterns which differ *intrinsically*, the overall set of physical *relations* which obtain among the activation patterns the networks generate in negotiating the task domain will be *common* across the various incarnations. Such relational physical structure reveals the kinds of contents conferred upon the networks' representing systems. Cluster analysis discloses the nature of the analogue, the topography of the representational landscape, that underwrites the computational performance of differing incarnations (at level-3) of the same type of network. So a tool such as cluster analysis affords a level-2 computational explanation of a cognitive capacity, since it concerns the disciplinary mechanism governing the semantically coherent processing of representation bearers. Those connectionists, such as Paul Churchland (1989b), who maintain that the correct level of explanation in connectionist cognitive science lies with behaviour of individual processing units are going to *miss* something *crucial* to an understanding of how PDP systems negotiate their task domains. They are left without an account of in what sense PDP networks engage in *computational* operations.

The upshot, then, of the investigation into both Horgan and Tienson's proposal and the computational operation of two actual PDP systems is that in connectionist cognitive science the question of content determination is not orthogonal to cognitive architecture. A (trained) PDP network is a physical system which embodies an analogue of a representational domain. When a physical engine incorporates an analogue of some aspect of the represented world, it has the capacity to behave in a semantically coherent manner, and thereby can be properly said to engage in computational operations. The structural isomorphism TCD, in short, renders connectionism a computational framework. As such, when it comes to the question of content determination, connectionists are committed to structural isomorphism.

Why do cognitive scientists (including Horgan and Tienson, who come so close) miss this result? We don't have to look far for an answer. Horgan and Tienson begin by simply assuming, as is standard in cognitive science, that the question of content determination is orthogonal to the issue of cognitive architecture. The "orthogonality assumption" – a vestige, so I've argued, of *classicism* – is so well-entrenched in cognitive science that when the idea of a semantic metric surfaces, cognitive scientists take content for granted and see themselves to be concerned with the *realisation* of content and the encoding of content relations in PDP systems. However, the strategy of assuming intentional content from the outset, and the subsequent focus on the realisation relation, paints a somewhat misleading picture of intentionality from a connectionist perspective. Within the PDP framework, representation bearers and the intentional states they comprise are the *result* of network training, the result of moulding a network's activation landscape. An activation landscape *becomes representational* in embodying an analogue of some aspect of the world. PDP systems do not possess, from the outset, a certain set of representation bearers poised for arrangement on an activation landscape such that their relative-position relations reflect their content relations. It's not that, prior to training, a network already has a stock of representation bearers, each with a certain profile and particular content, ready to realise intentional content and reflect content relations through being strategically placed on an activation landscape. Rather, representation bearers emerge as a result of a training regime. In the course of training a PDP system is taught to produce, out of all the possible activation patterns it could generate, a certain range of patterns in response to a certain array of inputs. The system's representation bearers are themselves being generated in the course of training. Moreover, the formation of representation bearers is *coextensive* with the determination of their content. Both coincide in the shaping of an activation landscape. That is, the emergence of a landscape topography that exhibits a structural isomorphism to a representational domain is simultaneously the formation of a network's representation bearers, and that which determines the content of such representation bearers. So, an answer to the question of content determination is not something to be contemplated once one already has an account of how semantic properties and relations are realised. Nor should we assume that the task is to attach the right contents to an established set of representation bearers. The activation patterns a network is disposed to generate, and the contents of its intentional states, are both a consequence of its training.

1.3.3. A Gradualist Perspective on Content Determination

The theory I've proposed above is in need of further reinforcement. I said that a PDP system's representational substrate, as a result of training, comes to embody an analogue of some aspect of the represented world, and in so doing the system acquires the capacity to represent that domain and engage in computational operations. Such a proposal could be seen to invite the following objection. "At what point in the process of a PDP network's modification do its representation bearers have their content conferred? The impression given by the above discussion is that content is conferred only once a network comes to manifest a structural isomorphism to its target domain. So, at some point (presumably at or near the completion of its training) a PDP system 'crosses the line' and its states suddenly become genuinely contentful. At some stage in the course of its learning a whistle blows, or whatever (to use a Fodorism), and, voila, we have content. This, however, is surely a rather implausible conception of how connectionist systems (or any cognitive systems, natural or artificial, for that matter) acquire content."

I agree that such a picture would constitute an implausible account of how information-processing systems come to possess content. And it is *not* the picture of content determination from the structural isomorphism perspective I wish to recommend. In approaching the close of this chapter's main argument, I want to sketch what I think is much closer to the true state of affairs regarding the emergence of a structural isomorphism and the determination of content.

The above objection operates with what might be described as a "total" and "completed" conception of an analogue. That is, at some stage an analogue becomes "locked-down", such that the structure of a representational substrate constitutes a veridical reflection of all relevant features of a target domain, and only at this point is representational content determined. This is an interpretation I wish to guard against. The development of an analogue is a gradual process, and the determination of content should be thought of in terms of a continuum. The process of training a network is one which gradually brings about increasingly definite and complex organisation to its representational substrate, and thereby content of greater dimensions and of an increasingly finer grain.

Prior to training, and even as training gets under way, a PDP system does not contain contentful entities (assuming, of course, a random assignment of weights). At this stage, the network is presented with a range of inputs, and responds by generating certain activation patterns which are all rather similar in profile. There is no systematic relationship between variation in a feature of the representational domain and variation in the profiles of activation patterns generated. Rather early on in a training regime, however, the determination of content begins. Gradually, the network generates a range of activation patterns which exhibit more defined variation from one another. Certain activation patterns, more similar in profile to one another than others, are produced in response to those inputs which are similar along certain dimensions. That is, there emerges a (somewhat crude, at this point), systematic relationship between the character of the inputs and the character of the activation patterns produced. In these early stages of training, a PDP network is engaged in a gross partitioning of its representational space in line with brute features of its representational domain. Its representational content is coarse-grained.

The process of training a network is one which brings about a *progressive refinement* of the relationship between the nature of the inputs and the nature of the activation patterns the network generates. The structure of the representing system becomes more complex, coming to reflect to a greater degree the structure of the representational domain. This increasingly complex organisation mirrors both a greater number of those dimensions which comprise the representational domain, and increasingly fine differences among the values along such dimensions. At the end of a training regime, a network possesses content of a finer grain than that had at the earlier stages of its training.

Activation landscape terminology can also be employed here. Early on in the training of a PDP network, its activation landscape is somewhat featureless, or homogeneous; the terrain consists of smooth, gently-rolling slopes. Over the course of training, the terrain undergoes alteration. As the network is taught to respond with a certain target activation pattern upon presentation of each input in the array, the activation landscape gradually changes, becoming increasingly delineated, or in technical terms, "bumpy".

So, content determination in PDP systems is a gradual and continuous process, one which takes place over the course of training, and proceeds from coarse-grained contents (at the early stages) through to comparatively fine-grained contents (as learning draws to a close). The way NETtalk's hidden-unit activation space is organised, or more precisely, the way such organisation develops over the course of the training regime, nicely illustrates the plausibility of this suggestion. NETtalk's ability to respond to different features of the representational domain develops at different rates. Sejnowski and Rosenberg (1987, 152) provide us with a glimpse into the order in which NETtalk learns to make useful discriminations along different features of the task domain. The network learns to discriminate between vowels and consonants at an early stage of its training. However, at this point it cannot distinguish between the individual vowels and individual consonants, and so its output consists of unintelligible babble. Thus far, NETtalk has effected only a crude partitioning of its hidden-unit activation space; the activation patterns it generates fall into either one of two major subvolumes of this space. Because of the network's "undeveloped" connection-weight configuration, activity propagated to the output layer results in outputs which display little differentiation. Another important stage occurs when NETtalk learns to respond to word boundaries. Once this is done, Sejnowski and Rosenberg report, NETtalk's output "resemble[s] pseudowords". Performance continues to improve over the course of training as the network makes more precise discriminations between the various letter-to-sound correspondences. Certain words can be discerned in the output stream, and eventually NETtalk's output is intelligible, exhibiting only a few errors. Interestingly, when NETtalk makes an error it typically produces a phoneme similar in sound to the correct one.

In general, then, it seems plausible to suppose that a PDP network learns to make useful discriminations along those coarser features of the representational domain prior to those more subtle, fine-grained, and contextual features. The gradual development of this more finely-tuned discriminative ability corresponds to the increasingly refined organisation of the network's representational substrate. A network's representational substrate takes on a more developed structure as training proceeds. The initially crude analogue becomes more definite and elaborate over the course of the learning period. As the system of representation bearers develops, that is, it exhibits a structural isomorphism to a wider range of features of the representational domain, as well as more fine-grained variation in each feature. Consequently, content becomes increasingly intricate through the course of the training regime.

A word of caution, however, is necessary here. What we regard – on the basis of a task analysis conducted *prior* to training – as salient dimensions of the task domain, and what we judge to be relatively coarse or fine-grained dimensions, may not correspond with those features important to the network's execution of its task. The PDP researcher's job is to unearth – via a post-training analysis – those features relevant to the network's negotiation of its task domain; those features, that is, by which the *network itself* defines the partitioning of its activation space (see Clark 1990).

The question of what features of the representational domain a given network is becoming "tuned to" (what features are responsible for a network's organisation of its hidden-unit activation space) is, in my view, a crucial one. And it is an area which PDP practitioners are probing with increasingly sophisticated tools of analysis. Less well understood is the closely related and equally important issue of the order in which a network learns to discriminate along these different dimensions. This is a matter which would obviously benefit from additional data on network organisation at various stages of a training regime. Nonetheless, I believe that there is already some reason for thinking that the "gradualist" position I've adopted is on the right track.

Finally, this is an appropriate point to take a last brief look at the "indeterminacy objection" to the structural isomorphism TCD. The centrepiece of my favoured response (first put forward in Chp. Four, sec. 3.), is that once we have reasonably complex, behaviourally versatile, and genuinely cognitive systems up and running – and thereby rather complex structural isomorphisms in place – it's just not likely that any representing system will be structurally isomorphic to a range of disparate representable domains. An analogue embodied by any particular neural representing system will be a highly complex and idiosyncratic structure. The PDP framework, I would now suggest, provides for the physical instantiation of the sophisticated kind of representing system posited by my favoured reply to the indeterminacy objection. A cognitively realistic PDP network consists of a *vast* number of processing units, all linked by finely and differentially weighted connection lines. The activation space of the hidden layer of such a network has, of course, as many dimensions as there are constituent processing units. And the particular (and massively complex) overall pattern of weighted connectivity determines the "shape" of such a hyper-dimensional activation space. The "deformations" created to this space as a result of a training regime constitute the moulding of a highly delineated representational landscape topography (see also Horgan 1997, 23). The nature of this landscape topography is so elaborate that, to my mind, the foundational premise of the indeterminacy objection is rather short-sighted. Indeed, to take NETtalk as an example, though the structural isomorphisms that I've claimed are implicated in PDP-style computation look to be there, they are so complex, so abstract, that they are difficult to clearly discern. And while NETtalk is meant to be suggestive of the *general* principles which may underlie our capacity to read aloud, it is hardly thought to be cognitively realistic to a significant extent. But even in the case of NETtalk, I would claim, there is simply *no domain other than English grapheme-to-phoneme conversion* to which the structure embodied in its representational substrate bears a structural isomorphism. As I also remarked in the earlier discussion, our understanding of an analogue has a long way to go to appreciate those which, in my view, Nature has employed in conferring content upon our internal representational states.

1.4. Securing Semantic Coherence: A Fundamental Difference Between Classicism and Connectionism

The position I've argued for above shows that the PDP framework presents a way of realising computational operations very different to that of its conventional rival. The literature contains numerous views on what marks connectionism's divergence from classicism, and one could be forgiven for regarding with suspicion (or perhaps fatigue) yet another candidate answer. However, the difference between classicist and connectionist computation I've sought to demonstrate over this section can with some justification lay claim to being described as a fundamental one. The reason is that it is a difference which goes to the *causal operation* of conventional and PDP devices *qua computational systems*. At bottom, the issue which has been the concern over section 1. is that of how *computational operations* are effected in conventional and PDP systems. *Qua* computational devices, what is it that drives the operation of conventional and PDP machines? Early in this section (subsec. 1.1.) I suggested that it is the notion of semantic coherence which is central to the nature of a computational operation. An account of the difference between classical and connectionist computation, then, must turn on the difference between how the competing computational frameworks are able to achieve semantically coherent processing. This, of course, is the significant difference I've argued for here. While the computational operations of conventional systems are determined by a set of state transition rules, PDP systems perform computational operations by virtue of utilising the structural isomorphism TCD. It is structural isomorphism that renders connectionism a *computational* framework, and marks it as a computational framework which is *distinct* from classicism.

Such a result provides an illuminating perspective on the proper level of analysis as regards the causal operation of conventional and PDP computational systems. A PDP system's representational substrate, I've argued, should be understood as embodying an analogue of a representational domain. The semantically coherent operation of such a system is determined by the physical profile of its representation bearers, in which, due to the presence of a structural isomorphism, their content and causal powers coincide. This is made possible because the physical constitution of a PDP system is such that its representational medium is directly fused with its representational substrate. Any variation in the physical property comprising the representational substrate is representationally, and thereby computationally, significant. The upshot is that the causal laws operative at the level of the system's representational substrate are those which govern its operation *qua computational device*. For this reason, PDP systems are analogue computers (see O'Brien 1998. O'Brien recasts the distinction between analogue and digital computation – typically drawn in terms of a continuous versus discrete representational medium, respectively – in terms of the *causal operation* of digital and analogue machines). An analogue computer embodies and utilises an analogue of a representational domain. The presence of an analogue means that the causal interactions among representation bearers are able to mirror the causal interactions among the representational objects such bearers stand-in for. The system's representational substrate is so configured that the causal laws which govern the representational substrate drive the system in a semantically coherent manner. The causal operation of a PDP computational system is to be explained in terms of the natural causal laws which apply to the physical property comprising the system's representational substrate.

Conventional computational systems are, of course, digital machines. The correct level of analysis as regards the causal operation of digital computers differs from that which applies to PDP systems. A digital computational system is created by first defining a representational medium over a representational substrate such that certain variations in the physical property comprising the representational substrate are not computationally significant. The representational medium of a digital computer is at one remove, or "divorced", from its representational substrate. The idea is to then drive the system not by those naturally occurring variations which exist at the level of the representational substrate, but by computational rules which are distinct from the substrate. Such rules are operative at the syntactic level. These rules apply, that is, to the *partitioned* values of the physical property which comprises the representational substrate. Such partitionings create the "transcendent" representational medium of a digital computer, and the system operates according to computational rules which apply to the syntactic properties of symbolic representation bearers incorporated in its representational medium. The computational operation of a digital computer is not directly governed by the natural causal laws which apply at the level of the system's representational substrate. Rather, the causal operation of a digital computer, *qua computational device*, is to be understood in terms of the computational rules operative at the level of its syntactic representational medium.[129]

Classicism and connectionism, as many cognitive scientists have thought, *are* different computational frameworks. What is perhaps surprising to many is that the important difference can be expressed in terms of differing implications for the naturalisation of intentionality project. It is standardly assumed that the question of content determination is independent of the nature of cognitive architecture. One's allegiance on matters of cognitive architecture, the orthodox thinking goes, has no bearing on one's choice of TCD. This assumption, I now hope to have shown, is an unexamined generalisation from classical cognitive science. While classical cognitive science is not bound to any particular TCD, connectionist cognitive science is committed to structural isomorphism.

2. Cognitive Science and the Naturalisation of Intentionality

The main argument of the current chapter is now in place. In the previous section I demonstrated that while classicists need not appeal to a TCD in their account of the mechanisation of rationality, connectionist computationalism implicates the structural isomorphism TCD. I propose to finish this chapter by reflecting on the differing circumstances classicists and connectionists find themselves in regard to the question of content determination. The moral of the following discussion will

[129] I am indebted to Gerard O'Brien (1993, personal communication) for the discussion of the digital and analogue nature of, respectively, conventional and PDP computational devices.

come as no surprise, and can be summed up as follows. Those who buy connectionism get a TCD included in the price, and if the arguments concerning the structural isomorphism TCD conducted herein are on track, then connectionists have happened upon a very good deal. On the other hand, stakeholders in classicism are free to go o'wanderin' in the TCD marketplace, but customers should be wary. As I've been concerned to show at various places in this work, it is not a buyer's market: the prices are high, and the goods are looking a little shop-soiled.

I'm going to begin in subsection 2.1. by examining the TCD options available to the classicist. It might be thought, however, that such an undertaking is largely misguided. After all, it was the burden of section 1. above to demonstrate that the prevailing orthogonality assumption is a product of classical cognitive science. So isn't it in some fashion illegitimate to consider possible classicism-TCD pairings? I have two closely related points to make in reply, two considerations against the view that the question of content determination can be addressed in complete isolation from cognitive science, classical cognitive science included.

First, the agenda of contemporary philosophy of mind is dominated by three main issues, "three great metaphysical puzzles": rationality, intentionality, and consciousness (Fodor 1991, 285). Of course, cognitive science – whether of classical or connectionist colours – is targeted at the rationality problem. But while the problem of explaining intentionality is a *distinct question* from that of explaining rationality, the prevailing strategy of addressing the naturalisation of intentionality in isolation from cognitive science is, in my view, past its use-by date. The reason is that the foundational claim of cognitive science has been enormously influential in recent thinking about the mind; it has influenced research in all those disciplines which investigate mentality and its physical instantiation (including, of course, philosophy of mind). After all, it was the rise of cognitive science which focused attention on the notion of mental representation, and provided the real momentum to the naturalisation of intentionality project. Indeed, as I remarked in Chapter One (subsec. 1.3.), the idea that cognitive processes are computational processes is the only workable explanation we currently have for how a purely physical system such as the brain could underwrite the rationality exhibited by complex cognitive systems. A science of the mind that is any way plausible cannot (at least at the current state of knowledge) do without the foundational claim of cognitive science. A *mature* science of the mind is aimed at resolving the three great puzzles of rationality, intentionality, and consciousness. And, of course, an answer provided for any one of these puzzles cannot be incompatible with an answer provided for another. Since the only problem about which there is broad consensus that real progress has been made is that of rationality (the explanation cognitive science provides), the naturalisation of intentionality cannot be conducted in complete isolation from cognitive science. A unified mature theory of mind – a theory, that is, which aims for a coherent, naturalistic explanation of rationality, intentionality, and consciousness – must be built on the *foundation* afforded by cognitive science.[130]

Second, and more stringently, a commitment to cognitive science is a commitment to mental representation, but the notion of mental representation cannot be left as an unexplained primitive. Cognitive science cannot make full use of the notion of mental representation in its account of the rationality of the cognitive mind, yet leave mental representation, so to speak, "unnaturalised". Again as indicated in Chapter One (subsec. 1.3.), cognitive science, whether it banks with classicism or connectionism, incurs an *explanatory debt* in purchasing the mechanisation of rationality on computational, and thereby representational, terms. The loan must be repaid in the currency of a naturalistic answer to the question of content determination. Only in this way can cognitive science lay claim to holding the controlling interest over a mature theory of mind. What I'm proposing to do in this section is, in effect, examine the differing repayment schedules confronting classicism and connectionism.

The naturalisation of intentionality, and the question of content determination in particular, cannot be addressed in isolation from cognitive science. So while the *classical* account of the semantically coherent processing of mental representation bearers doesn't implicate a TCD, it must still be explained how the contents of those "symbolic" items over which computational operations are defined are determined. This, of course, is something Jerry Fodor – perhaps the most prominent classicist, and arguably the most influential of contemporary theorists in setting the agenda for recent philosophy of mind – has long realised

[130] In subsection 2.2. I'm going to make more of the claim that cognitive science addresses the *foundational* issue in regard to the mind.

(see, in particular, his 1987). For some time Fodor has been concerned to wed his version of the causal TCD to his language of thought conception of cognitive architecture. Fodor has recognised, as classicists must, that the classical account of the mechanisation of rationality must be *augmented* with a naturalistic account of content determination.

2.1. Classicism and Content Determination

I now turn, then, to reflect on the situation classicists are in, given that they are free to choose from among those TsCD currently on offer. One would think that this would be an enviable position. However, I will show why this is not the happy state of affairs it might, *prima facie*, appear to be. This will not be a full-scale, intensive operation since it is secondary to the main line of argument already pursued in this chapter. Rather, I intend to drop-in on each of the possible classicism-TCD pairings and niggle a little.

The first point will take but a moment to state. I've already considered at some length the most significant hurdle confronting what I've been calling the "orthodox" TsCD. As I've demonstated (Chp. Two), all of the orthodox theories face serious difficulties in attempting to accommodate misrepresentation (see also Cummins 1996). The theorist who wants to couple classicism with one of the orthodox TsCD, then, is in the most uncomfortable position of having to wish very hard that, sooner rather than later, one or other of the standard proposals for dealing with misrepresentation can (in some amended fashion) be shown to work. But there is cause for considerable pessimism. As I've argued, it is the very *nature* of the orthodox approach to content determination and misrepresentation that is prohibiting a workable solution to the problem of misrepresentation.

In the earlier discussion of the orthodox theories, I did not explicitly consider functional role semantics (FRS), a TCD which has its fair share of adherents, especially amongst those of the classicist persuasion. Unfortunately for the classicist, however, FRS is not a satisfactory answer to the question of content determination. I'll mention three criticisms. First, as briefly mentioned above (see Chp. Four, note 97), FRS also operates within the orthodox content-determination framework, and so, as Cummins (1996) has argued, meets with precisely the difficulty of misrepresentation faced by the other orthodox TsCD. Second, as I argued above (Chp. Four, subsec. 1.4.), a functionalist TCD renders the explanatory role of content, at best, problematic, which is self-defeating. Third, Fodor (1990c, 428-429) has famously argued that FRS is seriously flawed as it leads to intentional irrealism (see also Fodor 1987, Chp. 3., and Fodor and Lepore 1992). Again, then, the classicist casting about for a way to discharge the explanatory debt of representationalism gets cold comfort from FRS.

I turn now to a further point concerning the TCD options available to the classicist. *Prima facie*, I've claimed, the classicist is free to choose from the orthodox TsCD, one such theory being teleosemantics. *Very* much *prima facie* as it turns out. A little reflection reveals that teleosemantics is not an attractive option for the classicist (or, indeed, the connectionist). Teleosemantics, it has been argued, is incompatible with computationalism in general. The reason, in outline, is that computationalism is committed to an *ahistorical* notion of content, and teleosemantics is essentially *historical* (see Cummins 1989, 80-81).

According to teleosemantics, content determination is essentially a matter of history. In short, a representation bearer has as its content that item it was selected to represent, that item the representation of which resulted in the representation bearer's replication. It is the selectional history of a representation-bearer type that determines its content. Consider now a molecule-for-molecule physical duplicate of a cognitive agent. *Ex hypothesi*, a physical duplicate of a cognitive agent has no selectional or learning history. So a consequence of the teleosemantics approach is that a physical duplicate does *not* have contentful internal states. But this conflicts with a basic tenet of computationalism, the commitment to an *ahistorical* understanding of content. According to computationalism, two computationally equivalent cognitive systems need not have common histories. Indeed, it's perfectly possible that the original cognitive subject and its duplicate, in being physically identical, have the same contentful internal states. Teleosemantics, therefore, is in significant tension with computationalism. Here's part of what Cummins has to say on the matter:

> ...the crucial issue is whether representation [that is, content] is grounded in the current state of the system, regardless of the history of that system. And on that issue the CTC [the computational theory of cognition] is absolutely unambiguous: Computationally equivalent

> states need share no historical properties at all. From a computational perspective, historical properties are accidental properties.
>
> ...
>
> It is which data structures [or, generically speaking, representation bearers] you have, not how you got them, that counts.... No account that (like Millikan's [teleosemantics]) takes the history of a data structure [or representation bearer] seriously can be right for the CTC (1989, 83-84).[131]

Teleosemantics, then, is a particularly salient example of the dangers of constructing a TCD in isolation from cognitive science. The fact that it looks to be incompatible (or at least in significant tension) with computationalism makes teleosemantics an unappealing option for the classicist.

Finally, there remains a rather glaring question. Are classicists able to advocate the *structural isomorphism TCD*? In my view, considerations to do with the conventional account of the physical realisation of computational processes indicate that the structural isomorphism TCD does not fit "naturally" with classical cognitive science. I must stress that the burden of the following discussion will not be to demonstrate that classicists are definitely prohibited from defending a version of structural isomorphism. Rather, I intend to worry about the ease with which classicism and the structural isomorphism TCD will come together. That is, I will be content with the weaker claim that it is at least not obvious how to hold both theories; structural isomorphism does not have a "natural affinity" with classicism.

On the structural isomorphism TCD, recall, variations in a physical feature of the represented system are mirrored by corresponding variations in a (non-equivalent) physical property of the represented system. Each representation bearer has its content by virtue of the isomorphism between the value-structure of the representing system and the value-structure of the represented system. This requires a representing system which admits of systematic physical variation. There must exist systematic physical relations among variations in a physical property of the representation bearers within the representing system. According to such a TCD, physical similarities between representation bearers constitute similarities in their content.

---

[131] Cummins' frames the point in terms of *classicism*. I intend it to apply to generic computationalism.

Consider now the conventional account of the physical realisation of a representational medium. A symbolic representational medium, as shown above, is one where some continuously variable physical property is systematically partitioned so as to generate *syntactic* structure. Such partitioning, that is, involves permitting a certain range of variation in the physical property without a difference in syntax. Under interpretation, the partitioned physical property becomes a symbolic representational medium. And because it is onto the syntactic properties of representation bearers that their content is mapped, certain physical variations in tokens of the one type will not constitute a difference in content. I can say, then, that the representational *medium* of a classical system is not directly "fused" with its representational *substrate* (the latter being, as pointed out above, the physical property which implements the former). In other words, the representational mediuin exists, as it were, at one remove from the physical property which ultimately implements the symbolic representation bearers, such that the syntactic structure of the representational medium is not directly that of the "natural" physical structure exhibited by the representational substrate. The structure exhibited by a symbolic representational medium is, so to speak, imposed from the outside, in that it must be understood relative to the system's state transition rules. Consequently, the relations which obtain among symbolic representation bearers are not to be understood as existing at the "level" of the system's representational substrate. Such relations are, from the point of view of the physical property which constitutes the representational substrate, somewhat arbitrary or artificial. The relations symbols bear to one another stem from the syntactic structure imposed upon the implementing substrate.

So, the relations symbols bear to one another are syntactic rather than straightforwardly physical. And it is onto such syntactic properties that external objects are mapped. But the structural isomorphism TCD requires a mapping of the value-structure of the represented system onto the *physical* structure of the representing system. It therefore looks as if the structural isomorphism TCD does not find a natural home with classicism.

Consider again the structural isomorphism TCD. The representation bearers of the representing system represent objects in the represented system because there is an isomorphism between the physical relations among representation bearers and certain physical relations among their objects. Two representation bearer tokens have the same content when they have the same set of physical relations to the other representation bearers in the system. Similarities and differences in representation bearer content is a matter of similarities and differences in a physical property of representation bearers.

As indicated, in a classical system content is mapped onto the syntactic properties of representation bearers, and the representation bearers bear syntactic relations to one another. These syntactic properties must be understood relative to the *rules* which govern the system's state transitions. The rules determine the pattern of actual and counterfactual causal interactions into which symbols enter. This means that two tokens are of the same syntactic type when they have the same causal role. And the relations symbols bear to one another are in terms of their respective causal roles. So, an isomorphism between the relations among symbols and the relations among the representaional objects would be between the *causal* network of symbols and the network of representational objects. This, however, is a *functional*, not structural, isomorphism. The notion of functional isomorphism, as indicated (subsec. 1.1.), is another way of understanding the semantically coherent processing of representation bearers. It is not surprising, then, that many classicists endorse functional role semantics. Such a view of content determination is the idea that the functional isomorphism which constitutes semantically coherent processing in classical systems is that which determines the content of the symbolic entities such systems manipulate. Functional role semantics is naturally compatible with classicism. The same cannot be said, I think, for structural isomorphism.

Content determination is not bound-up with the conventional means of disciplining transactions among representation bearers. The classicist, in terms of the task of attaining semantically coherent processing, is not forced to embrace one particular TCD. Now, as every consumer in a capitalist society knows, choice is a Good Thing. However, as every voter in a democratic society knows, a choice from a range of unpalatable options is not so much fun. This is the kind of situation, I've argued here, in which classicists find themselves. The classicist confronting the question of content determination is faced with a choice between a number of unpalatable TsCD.

2.2. Computation, Representation, and Connectionism

So far in this section I've been concerned with the idea that the naturalisation of intetionality cannot be conducted in isolation from a computational account of cognitive processes. Such a claim turns on considerations pertaining to the aspirations of a mature theory of mind, and, relatedly, the explanatory debt incurred by computationalism in employing representation in its solution to the rationality problem. The considerations rehearsed above provide for a certain, "minimal" relationship between cognitive science and the naturalisation of intentionality: the problem of intentionality is to be addressed by building on the groundwork supplied by a computational understanding of cognitive processes. A fuller account of cognition of a whole (both cognitive processing and the contents of representation bearers whose manipulations constitute cognitive operations) requires that a computational framework be supplemented with a TCD.

I'm now going to put forward a reason for supposing that the relationship between cognitive science and the naturalisation of intentionality is *tighter* than outlined above. The usual view is that the great metaphysical puzzles of rationality and representation are issues of equal standing which sit side by side, related only in that minimal sense which Jerry Fodor, as pointed out above, has done much to reinforce. On the line of thinking I'm now going to develop, representation is not an issue of independent standing to that of rationality, rather representation and computation are *deeply intertwined.* Representation is founded on computation. This way of interpreting the relationship between cognitive science and the naturalisation of intentionality has content determination as a question which must be addressed *from within cognitive science itself*, rather than simply an issue which must be broached to augment a computational theory of cognitive processes. Given the argument of section 1. above, the upshot is that only *connectionism* – in having semantic coherence and content determination bound-up together in the form of structural isomorphism – provides a robust account of the idea that computation and representation are, for reasons more general than the details of a particular computational framework, fundamentally intertwined. Connectionism, but not classicism, can make good on the claim that it's the business of cognitive science to say how mental content is conferred.[132]

In my view, there is a general, principled reason why the question of content determination must be the business of cognitive science. As hinted above, it has to do with the "status" of each of the three great metaphysical puzzles about the mind. One gets the impression that contemporary philosophers think the problems of rationality, representation, and consciousness are created equal. To my mind, this is not right. The problem of rationality, the issue of what underwrites our capacity for intelligent behaviour, is *the* great mystery of the mind. *It* is the problem Nature confronted. And Nature solved it by constructing *computational* systems. *Computation is the foundational notion.* Minds are representational systems only insofar as they are computational systems. Cognitive science, then, sits at the foundation of an overall, naturalistic account of the mind.

The mistake of the prevailing view is that it fails to appreciate the intimate connection between representation and computation. The standard interpretation, that is, misconceives the role of representation in the emergence of rational physical systems. The problem of representation is not on a par with the problem of rationality. Representational creatures only arose out of the need to construct computational systems. The problem Nature confronted was that of building *intelligent* systems (not intentional systems, or conscious systems). The capacity for intelligent behaviour is the primary issue. Natural selection constructed computational devices because computational devices are rational or semantically coherent physical systems, those which give rise to intelligent behaviour. And computational systems require a method of realising information or representational content, a collection of representing systems. Intentionality arose only as a *means* to creating computational systems, a means to effecting intelligent behaviour. The evolution of systems with intentional states would have been of no use (Nature would not have "designed" or selected organisms with internal representational states) if such states did not have content-relevant effects on other internal states of

[132] In setting out to argue that an answer to the question of content determination must come from within cognitive science, I am not denying that the problem of rationality and the problem of mental representation are, strictly speaking, separate issues. The question of how to naturalistically explain the causal processes which are responsible for our intelligent behaviour, and the question of how to naturalistically explain mental content determination, are distinct questions. I am not suggesting that an "unadorned" answer to the rationality problem *ipso facto* yields a solution to the problem of content determination. This would be to commit the kind of fundamental error present in Searle's (1981) "Chinese Room" argument.

the organism, and, ultimately, on its behavioural responses. Being a content-bearing internal item with content-relevant effects (having content-relevant causal powers) just *is* the notion of a computation. So, the problem Nature initially solved was that of rationality. Computation is explanatorily prior to representation, such that the question of content determination arises *inside* the question of rationality. And the computational theory of mind is the only route to an *explanation* of rationality. Therefore an answer to the question of content determination must come from *within* cognitive science.[133]

The view of connectionism I've argued for is, then, elegantly consistent with what I think is the correct interpretation of the relationship between the problems of rationality and representation. And so the main argument of this chapter – namely, that the connectionist account of semantically coherent processing is, in contrast to the classical account, bound-up with a TCD – is not, as it were, a "neutral" result. There is reason for thinking that the connectionist approach is to be preferred. More broadly, given the arguments concerning the structural isomorphism TCD offered in this work, the fact that connectionism implicates structural isomorphism means that connectionists have inherited a TCD with a number of significant advantages over the orthodox theories.

Connectionism, articulated correctly, provides an explanation of cognition as a whole where the demands of securing semantic coherence and the determination of content are simultaneously satisfied. There is an element of *efficiency* in the connectionist account which strikes me as attractive. Contrast this with classicism, where semantic coherence and content determination are distinct issues and distinct operations. With the account of semantic coherence in place, the classicist *then* has to cast about for a suitable TCD. The classical theorist engages in a "two-stage" explanation of cognition as a whole. I'm not suggesting that this

[133] The same line of argument can be employed to show that an explanation of consciousness must also come from within cognitive science (given the plausible intuition that consciousness must have emerged to *do something*, that is, have causal impact upon an organism's cognitive economy). Nature employed representation in her computational solution to the problem of intelligence, and consciousness as her means of further *finessing* that solution.

counts as a decisive objection to classicism, but it is, I think, a more involved task when compared to the economical alternative offered by connectionism.[134]

The real significance of the emergence of connectionism – the important departure from the approach of traditional cognitive science – is that it forces a quite dramatic shift in attention onto the material substrate of cognitive agents, in particular, the representation bearers of cognition. Under the dominance of classicism, cognitive science has been conducted in relative isolation from details of the physical properties of representation bearers. The focus has been on cognitive operations, the processes by which transformations of mental representation bearers take place. Of course, symbolic representation bearers are a *part* of the classical explanation of cognition. But symbols do not automatically engage in content-appropriate transactions. Symbol manipulations must be constrained by computational rules, which apply not to the intrinsic properties of symbols but to their syntactic properties; such higher-order physical properties must therefore be understood in relation to the rules which govern symbol transformations. Moreover, the intrinsic properties of symbols are not thought to be relevant to the contents they possess. There is an arbitrary relationship between the intrinsic physical properties of symbols and their contents. Rather, what matters is (typically) the externally-instituted relation between representational objects and the tokening of internal symbolic items. On the connectionist account of cognition, in contrast, mental representation bearers – neural representing systems – assume a vitally important role. For the connectionist, it is the *causal powers of the mental representation bearers themselves* which are crucial to an explanation of cognition. The representation bearers themselves effect content-appropriate transactions in terms of their particular physical profiles. By virtue of belonging to a system of

[134] Actually, I inclined to think that there is not just *theoretical* efficiency in having semantically coherent processing and content determination bound together in connectionist cognitive science, but also *biological* or *evolutionary* efficiency. The move from the theoretical to the biological claim requires more work than I have space for here, so let me just sketch the picture with broad brushstrokes. Suppose, for a moment, that the main argument of this chapter is correct, and that, moreover, Mother Nature is a connectionist. So the mind has a connectionist architecture, and structural isomorphism is the *mechanism* Mother Nature employed for achieving semantic coherence and the conferral of content. In thinking along very general evolutionary lines, there is an air of efficiency in having the demands of securing semantic coherence and content determination simultaneously satisfied. Mother Nature, we have come to realise, is something of a miser (a "satisficer", to use the famous term), prone to making-do with the least costly means of meeting certain ends. A *mechanism* by which semantics and semantic coherence are handled *together* is just the sort of mechanism we would expect Mother Nature to "design". Such a claim is particularly salient if, as I've argued in this subsection, computation and representation are inextricably tied together in Mother Nature's solution to the problem of rationality.

representation bearers – a system whose relational physical structure manifests a structural isomorphism to an aspect of its representational domain – each representation bearer's physical profile embodies its content-appropriate causal powers.

So, the connectionist account of cognition focuses attention on the physical properties of representation bearers. Semantically coherent processing is effected by the employment of a structural isomorphism between a system of representation bearers and a represented system. The structural isomorphism TCD says it is the physical properties of representation bearers that determine their content. An account of semantic coherence and a theory of content determination thereby coincide, for connectionism, in the representation bearers of cognition.

CONCLUSION

In this work, I've sought to question and supplant some traditional assumptions about mental representation. Most centrally, the structural isomorphism TCD, over the range of issues considered, articulates a perspective on mental representation firmly grounded in the nature of mental representation bearers themselves. Structural isomorphism thereby offers an account of content determination which departs significantly from the orthodox approach. Rather than content determination being a matter of the world itself impinging upon the brain, I've argued that we have to look at content determination from precisely the other way around. It is the relational physical structure of a neural representing system which determines the contents of its constituent mental representation bearers. Such relational physical structure is constituted by the systematic relationships between the intrinsic physical properties of representation bearers. By virtue of its relational physical structure, a representing system itself determines its represented system.

In according the primary content-determining role to the physical properties of mental representation bearers themselves, I've done something which is singularly unpopular in contemporary philosophy of mind. It is standardly assumed that the physical properties of mental representation bearers are not crucial, indeed, not significant, to content determination, and, moreover, to cognition as a whole. I've attempted to demonstrate herein that it's time to think otherwise. The structural isomorphism TCD possesses a number of substantial advantages. It constitutes a use-independent TCD which operates within an alternative approach to misrepresentation, one that, I've argued, overcomes the problem of misrepresentation which has so bedeviled the orthodox theories. It also effortlessly handles both the robustness of content and what I've termed the "selectivity" of mental representation, the latter a phenomenon which has been illegitimately neglected in contemporary discussions. The structural isomorphism TCD I've formulated operates with a fine-grain conception of content – "microcontent" – one in keeping, so I've argued, with both the neurophysiological evidence and theoretical and empirical considerations concerning conscious representational experience. The proprietary, physically structured representing systems of structural isomorphism naturally account for the structured character of representational experience. I've shown that while certain empirical findings from the representational domains of temperature and colour serve to seriously undermine the prevailing externalist approach to content determination, such results are strongly suggestive of the kind of internalist TCD structural isomorphism provides. I've also sought to build a bridge between cognitive science and the concerns of contemporary philosophy of mind. I've argued that it is structural isomorphism which secures the semantically coherent processing of representation bearers in PDP systems. Content determination is not orthogonal to cognitive architecture in connectionist cognitive science. Connectionism is committed to the structural isomorphism TCD. Connectionism offers a theory of cognition as a whole whereby the content-appropriate causal powers of representation bearers reside in their physical properties. Though I stand by the main argument of Chapter Six, I do have a *caveat* at this point. The structural isomorphism TCD does not stand or fall with connectionism. It is, I think, independently viable. Importantly, however, connectionism does provide *a* physical *instantiation* of the structural isomorphism TCD.

The explanation of content determination offered by structural isomorphism does, however, bring with it a significant re-conception of mental content. I first investigated an alternative understanding of intentionality independently of structural isomorphism. Only with the discussion of Chapter Five, however, does the radical departure of the conception of content proposed herein from the orthodox conception become particularly stark. The world of mental content, I've argued, is not straightforwardly the world itself. Rather, it is the world-*as-represented*, a *notional* world, a "cleaned-up" external world. No one ever thought that understanding mental representation would be easy. The advent of the notional world shows just how intriguing and remarkable the property of intentionality is, just how deftly Nature conferred representational capacities upon her creatures.

Of course, this work only *begins* the development and defence of the structural isomorphism TCD. There is much here that would clearly benefit from further treatment, a point made salient with just few examples. It would, for instance, be profitable to enter the structural isomorphism TCD into the established and ongoing debate over the causal role of content. The previously unexplored area of intentional object fixation (Cummins 1996) has recently emerged as a new issue.

The very idea of a notional world is novel and difficult, and, I suppose, one of the more radical and speculative hypotheses proposed herein. It requires further articulation and defence. A detailed investigation into the operation of a number of representing systems across the various modalities may bring rewards here. And, of course, the claim that connectionism implicates structural isomorphism will only be fully substantiated with an examination into the organisation of many more actual PDP systems.

As regards the structural isomorphism TCD in general, it might be objected that it won't "scale-up" to account for those more sophisticated representational capacities we possess. I have three points in reply. First, I signaled from the early stages of this work (Chp. One, subsec. 1.4.) that I was explicitly targeting those more basic domains of representational activity, the "representational base". Strictly speaking, then, the above charge is illegitimate. Second, every other TCD is in the same boat here. It's not obviously apparent how the causal and teleological theories will scale-up to higher cognition. The above objection, then, is not one that mitigates against structural isomorphism in favour of the orthodox theories. Indeed, I think the development of the structural isomorphism TCD conducted over this work is in fact founded on more cognitively realistic examples than is usual, and so the objection applies with *less* force to structural isomorphism than existing accounts of content determination. Dretske and Millikan, for instance, seek illustration and support from very simple, non-human representational capabilities, and even non-mental examples. In contrast, I've derived many of the claims herein from empirical findings on the constitution and operation of sensory systems in the *human* case, and, in the discussion of connectionism, from suggestive modeling of a decidedly human capability, namely, text to speech conversion. This is much better than thermometers, marine bacteria, bee dances, and frogs poking their tongues out at flies. Third, I suspect that an unimaginative conception of an analogue is responsible for much of the intuitive force of the objection at hand. As I've highlighted on a number of occasions above, the analogues purportedly involved in mental representation will have a highly complex and abstract nature. It's early days in the development of the structural isomorphism TCD, and the tantalising suggestion is that we are seeing only the beginnings of our conception of an analogue. As connectionism taught us with regard to the idea of superpositional encoding, Nature may well do things in a manner we had never previously imagined.

Whatever the ultimate verdict on the TCD developed in this work, and the other central ideas proposed herein, I do finally want to highlight something not explicitly indicated beforehand. I see much of the present work as a kind of "conceptual underlabouring" of the very notion of mental representation. I've sought here to bring into the open a number of distinctions concerning mental representation which have previously been either unrecognised or insufficiently appreciated.

I began this work (in the Preface) with a game metaphor, so it's only appropriate that I end with one. My overall aim in this work has been to show that when it comes to the question of content determination there is, in the shape of structural isomorphism, a new game in town. And if the arguments conducted herein are anywhere near on-track, it's a game that an increasing number of philosophers are going to want to play.

BIBLIOGRAPHY

Akins, K. (1996). 'Of Sensory Systems and the "Aboutness" of Mental States.' *Journal of Philosophy* 93: 337-372.

Armstrong, D. (1968). *A Materialist Theory of the Mind.* (London, Routledge and Kegan Paul).

Armstrong, D. (1978). *A Theory of Universals.* Vol. II of *Universals and Scientific Realism.* (Cambridge, Cambridge University Press).

Baars, B. (1988). *A Cognitive Theory of Consciousness.* (Cambridge, Cambridge University Press).

Baker, L. (1989). 'Instrumental Intentionality.' *Philosophy of Science* 56: 303-316.

Bechtel, W. (1988). 'Connectionism and Interlevel Relations.' *Behavioral and Brain Sciences* 11: 24-25.

Bechtel, W. (1998). 'Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science.' *Cognitive Science* 22: 295-318.

Bechtel, W. and Abrahamsen, A. (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks.* (Cambridge, Mass., Blackwell).

Beer, R. (1990). *Intelligence and Adaptive Behavior.* (San Diego, Academic Press).

Blinder, D. (1986). 'In Defense of Pictorial Mimesis.' *Journal of Aesthetics and Art Criticism* 45: 19-27.

Block, N. (1986). 'Advertisement for a Semantics for Psychology.' In P. French, T. Uehling Jr., and H. Wettstein (eds.), *Midwest Studies in Philosophy: Studies in the Philosophy of Mind.* Vol. 10. (Minnesota, University of Minnesota Press).

Block, N. (1990). 'The Computer Model of the Mind.' In D. Osherson and E. Smith (eds.), *Thinking: An Invitation to Cognitive Science.* Vol. 3, 247-289. (Cambridge, Mass., MIT Press/Bradford).

Boghossian, P. and Velleman, D. (1989). 'Colour as a Secondary Quality.' *Mind* XCVIII: 81-103.

Boynton, R. (1979). *Human Color Vision* (New York, Holt, Rinehart, and Winston).

Brentano, F. (1874/1973). *Psychology From an Empirical Standpoint*. L. McAlister (ed.) (London, Routledge and Kegan Paul).

Brooks, R. (1991). 'Intelligence Without Representation.' *Artificial Intelligence* 47: 139-159.

Churchland, P. M. (1981). 'Eliminative Materialism and the Propositional Attitudes.' *Journal of Philosophy* 78: 67-90.

Churchland, P. M. (1988). *Matter and Consciousness*. (Cambridge, Mass., MIT Press/Bradford).

Churchland, P. M. (1989a). 'Some Reductive Strategies in Cognitive Neurobiology.' In his *A Neurocomputational Perspective*. (Cambridge, Mass., MIT Press/Bradford).

Churchland, P. M. (1989b). 'On the Nature of Theories: A Neurocomputational Perspective.' In his *A Neurocomputational Perspective*. (Cambridge, Mass., MIT Press/Bradford).

Churchland, P. M. (1990). 'Cognitive Activity in Artificial Neural Networks.' In D. Osherson and E. Smith (eds.), *Thinking: An Invitation to Cognitive Science*. Vol. 3, 199-227. (Cambridge, Mass., MIT Press/Bradford).

Churchland, P. M. (1995). *The Engine of Reason, The Seat of the Soul*. (Cambridge, Mass., MIT Press/Bradford).

Churchland, P. S. and Sejnowski, T. (1989). 'Neural Representation and Neural Computation.' In L. Nadel, L. Cooper, P. Culicover, and R. Harnish (eds.), *Neural Connections, Mental Computation*. (Cambridge, Mass., MIT Press/Bradford).

Churchland, P. S. and Sejnowski, T. (1992). *The Computational Brain*. (Cambridge, Mass., MIT Press/Bradford).

Clark, Andy (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. (Cambridge, Mass., MIT Press/Bradford).

Clark, Andy (1990). 'Connectionism, Competence, and Explanation.' In M. Boden (ed.), *The Philosophy of Artificial Intelligence*. (Oxford, Oxford University Press).

Clark, Andy (1993a). *Associative Engines: Connectionism, Concepts, and Representational Change*. (Cambridge, Mass., MIT Press/Bradford).

Clark, Andy (1993b). 'Minimal Rationalism.' *Mind* 102: 587-610.

Clark, Andy (1998). 'Time and Mind.' *Journal of Philosophy* 95: 354-376.

Clark, Andy and Toribio, J. (1994). 'Doing Without Representing?' *Synthese* 101: 401-431.

Clark, Austen (1993). *Sensory Qualities.* (Oxford, Clarendon Press).

Cornsweet, T. (1970). *Visual Perception.* (New York, Academic Press).

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction.* (Cambridge, Mass., Blackwell).

Craik, K. (1943). *The Nature of Explanation.* (Cambridge, Cambridge University Press).

Crane, T. (1995). *The Mechanical Mind* (London, Penguin).

Cummins, R. (1989). *Meaning and Mental Representation.* (Cambridge, Mass., MIT Press/Bradford).

Cummins, R. (1996). *Representations, Targets, and Attitudes.* (Cambridge, Mass., MIT Press/Bradford).

Cummins, R. and Schwarz, G. (1991). 'Connectionism, Computation, and Cognition.' In T. Horgan and J. Tienson (eds.), *Connectionism and the Philosophy of Mind.* (Dordrecht, Kluwer).

Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology.* (Montgomery, VT, MIT Press/Bradford).

Dennett, D. (1987). *The Intentional Stance.* (Cambridge, Mass., MIT Press/Bradford).

Dennett, D. (1991a). 'Real Patterns.' *Journal of Philosophy* 89: 27-51.

Dennett, D. (1991b). *Consciousness Explained.* (Boston, Little, Brown, and Company).

Dietrich, E. (1989). 'Semantics and the Computational Paradigm in Cognitive Psychology.' *Synthese* 79: 119-141.

Dretske, F. (1981). *Knowledge and the Flow of Information.* (Oxford, Blackwell).

Dretske, F. (1986/1990a). 'Misrepresentation.' In W. Lycan (ed.), *Mind and Cognition.* (Cambridge, Mass., MIT Press/Bradford).

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes.* (Cambridge, Mass., MIT Press/Bradford).

Dretske, F. (1990b). 'Does Meaning Matter?' In E. Villanueva (ed.), *Information, Semantics, and Epistemology.* (Oxford, Blackwell).

Dretske, F. (1993). 'Conscious Experience.' *Mind* 102: 263-283.

Dretske, F. (1995). *Naturalizing the Mind.* (Cambridge, Mass., MIT Press/Bradford).

Edelman, G. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection.* (New York, Basic Books).

Field, H. (1977). 'Logic, Meaning, and Conceptual Role.' *Journal of Philosophy* 74: 379-409.

Field, H. (1978). 'Mental Representation.' *Erkenntnis* 13: 9-61.

Files, C. (1996). 'Goodman's Rejection of Resemblance.' *British Journal of Aesthetics* 36: 398-412.

Fodor, J. (1975). *The Language of Thought.* (Cambridge, Mass., MIT Press/Bradford).

Fodor, J. (1980). 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology.' *Behavioral and Brain Sciences* 3: 63-110.

Fodor, J. (1984/1990a). 'Semantics, Wisconsin Style.' In his *A Theory of Content and Other Essays.* (Cambridge, Mass., MIT Press/Bradford).

Fodor, J. (1985/1990a). 'Fodor's Guide to Mental Representation.' In his *A Theory of Content and Other Essays.*

Fodor, J. (1986). 'Why Paramecia Don't Have Mental Representations.' In P. French, T. Uehling Jr., and H. Wettstein (eds.), *Midwest Studies in Philosophy: Studies in the Philosophy of Mind.* Vol. 10. (Minnesota, University of Minnesota Press).

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* (Cambridge, Mass., MIT Press/Bradford).

Fodor, J. (1990a). *A Theory of Content and Other Essays.* (Cambridge, Mass., MIT Press/Bradford).

Fodor, J. (1990b). 'Reply to Dretske's "Does Meaning Matter?"' In E. Villanueva (ed.), *Information, Semantics, and Epistemology.* (Oxford, Blackwell).

Fodor, J. (1990c). 'Banish Discontent.' In W. Lycan (ed.), *Mind and Cognition.* (Cambridge, Mass., MIT Press/Bradford).

Fodor, J. (1991). 'Replies.' In B. Loewer and G. Rey (eds.), *Meaning in Mind: Fodor and his Critics* (Oxford, Blackwell).

Fodor, J. (1992). 'The Big Idea: Can There be a Science of the Mind?' *Times Literary Supplement* July 3: 5-7.

Fodor, J. (1994). *The Elm and the Expert: Mentalese and its Semantics.* (Cambridge, Mass., MIT Press/Bradford).

Fodor, J. and Pylyshyn, Z. (1988). 'Connectionism and Cognitive Architecture: A Critical Analysis.' *Cognition* 35: 3-71.

Fodor, J. and McLaughlin, B. (1990). 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work.' *Cognition* 35: 183-204.

Fodor, J. and Lepore, E. (1992). *Holism: A Shopper's Guide.* (Oxford, Blackwell).

Fodor, J. and Lepore, E. (1996). 'Paul Churchland and State Space Semantics.' In R. McCauley (ed.), *The Churchlands and their Critics.* (Cambridge, Mass., MIT Press/Bradford).

Gardenfors, P. (1996). 'Mental Representation, Conceptual Spaces, and Metaphors.' *Synthese* 106: 21-47.

Gilman, D. (1972). 'A New Perspective on Pictorial Representation.' *Australasian Journal of Philosophy* 70: 174-186.

Godfrey-Smith, P. (1989). 'Misinformation.' *Canadian Journal of Philosophy* 19: 533-550.

Godfrey-Smith, P. (1992). 'Indication and Adaptation.' *Synthese* 92: 283-312.

Goodman, N. (1969). *Languages of Art.* (London, Oxford University Press).

Hardin, C. (1984a). 'Are 'Scientific' Objects Coloured?' *Mind* XCIII: 491-500.

Hardin, C. (1984b). 'A New Look at Color.' *American Philosophical Quarterly* 21: 125-133.

Hardin, C. (1985). 'The Resemblances of Colors.' *Philosophical Studies* 48: 35-47.

Hardin, C. (1988). *Color for Philosophers.* (Indianapolis/Cambridge, Hackett).

Hardin, C. (1990). 'Color and Illusion.' In W. Lycan (ed.), *Mind and Cognition* (Cambridge, Mass., MIT Press/Bradford).

Harman, G. (1982). 'Conceptual Role Semantics.' *Notre Dame Journal of Formal Logic* 23: 242-256.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* (Cambridge, Mass., MIT Press).

Haugeland, J. (1987). 'An Overview of the Frame Problem.' In Z. Pylyshyn (ed.), *The Robot's Dilemma.* (Norwood, New Jersey, Ablex).

Hensel, H. (1982). *Thermal Sensations and Thermoreceptors in Man.* (Springfield, Thomas).

Hilbert, D. (1987). *Color and Color Perception: A Study in Anthropocentric Realism*. (Chicago, University of Chicago Press).

Horgan, T. (1997). 'Connectionism and the Philosophical Foundations of Cognitive Science.' *Metaphilosophy* 28: 1-30.

Horgan, T. and Tienson, J. (1989). 'Representations Without Rules.' *Philosophical Topics* 17: 27-43.

Horgan, T. and Tienson, J. (1996). *Connectionism and the Philosophy of Psychology*. (Cambrige, Mass., MIT Press/Bradford).

Horgan, T. and Woodward, J. (1985). 'Folk Psychology is Here to Stay.' *Philosophical Review* 94: 197-226.

Hubel, D. (1988). *Eye, Brain, and Vision*. (New York, Freeman).

Hurvich, L (1981). *Color Vision*. (Sunderland, Mass., Sinauer).

Hurvich, L (1985). 'Opponent-colours Theory.' In D. Ottoson and S. Zeki (eds.) *Central and Peripheral Mechanisms of Colour Vision*. (London, MacMillan).

Lloyd, D. (1988). 'Connectionism in the Golden Age of Neuroscience.' *Behavioral and Brain Sciences* 11: 42-43.

Lloyd, D. (1989). *Simple Minds* (Cambridge, Mass., MIT Press/Bradford).

Lloyd, D. (1995). 'Consciousness: A Connectionist Manifesto.' *Minds and Machines* 5: 161-185.

Lloyd, D. (1996). 'Consciousness, Connectionism, and Cognitive Neuroscience: A Meeting of the Minds.' *Philosophical Psychology* 9: 61-79.

Loar, B. (1981). *Mind and Meaning*. (Cambridge, Cambridge University Press).

Loar, B. (1982). 'Conceptual Role and Truth Conditions.' *Notre Dame Journal of Formal Logic* 23: 272-283.

Maloney, C. (1990). 'Mental Misrepresentation.' *Philosophy of Science* 57: 445-458.

Marr, D. (1982). *Vision*. (New York, Freeman).

Marr, D. (1990). 'Artificial Intelligence: A Personal View.' In M. Boden (ed.), *The Philosophy of Artificial Intelligence*. (Oxford, Oxford University Press).

Manfredi, P. and Summerfield, D. (1992). 'Robustness Without Asymmetry: A Flaw in Fodor's Theory of Content.' *Philosophical Studies* 66: 261-283.

McClelland, J. and Rumelhart, D. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 1: Foundations.* (Cambridge, Mass., MIT Press).

McClelland, J. and Rumelhart, D. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 2: Psychological and Biological Models.* (Cambridge, Mass., MIT Press).

McGinn, C. (1989). *Mental Content.* (Oxford, Blackwell).

Millikan, R. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism.* (Cambridge, Mass., MIT Press/Bradford).

Millikan, R. (1993). *White Queen Psychology and Other Essays for Alice.* (Cambridge, Mass., MIT Press/Bradford).

Nagel, T. (1974). 'What is it Like to be a Bat?' *Philosophical Review* 83: 435-450.

Neander, K. (1987). 'Pictorial Representation: A Matter of Resemblance.' *British Journal of Aesthetics* 27: 213-226.

Neander, K. (1991a). 'Functions as Selected Effects: The Conceptual Analyst's Defence.' *Philosophy of Science* 58:

Neander, K. (1991b). 'The Teleological Notion of "Function".' *Australasian Journal of Philosophy* 69:

O'Brien, G. (1993). *The Computational Theory of Mind: An Exploration in the Conceptual Foundations of Cognitive Science.* (Unpublished doctoral dissertation, Oxford University).

O'Brien, G. (1998). 'Connectionism, Analogicity, and Mental Content.' *Acta Analytica* 22: 111-131.

O'Brien, G. and Opie, J. (1998). 'The Disunity of Consciousness.' *Australasian Journal of Philosophy* 76: 378-395.

O'Brien, G. and Opie, J. (1999a). 'A Connectionist Theory of Phenomenal Experience.' *Behavioral and Brain Sciences* 22: 1-25.

O'Brien, G. and Opie, J. (1999b). 'Putting Content into a Vehicle Theory of Consciousness.' (Author's response to peer commentary). *Behavioral and Brain Sciences* 22: 50-72.

Palmer, S. (1978). 'Fundamental Aspects of Cognitive Representation.' In E. Rosch and B. Lloyd (eds.) *Cognition and Categorization.* (Hillsdale, Lawrence Erlbaum).

Papineau, D. (1987). *Reality and Representation.* (Oxford, Blackwell).

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

Peirce, C. (1931-1958). *Collected Papers of Charles Sanders Peirce.* C. Hartshorne, P. Weiss, and A. Burks (eds.) (Cambridge, Mass., Harvard University Press).

Pylyshyn, Z. (1984). *Computation and Cognition.* (Cambridge, Mass., MIT Press).

Poole, D. (1974). 'Goodman and the "Naïve" View of Representation.' *British Journal of Aesthetics* 14: 68-80.

Sartwell, C. (1994). 'What Pictorial Realism Is.' *British Journal of Aesthetics* 34: 2-12.

Searle, J. (1981). 'Minds, Brains, and Programs.' *Behavioral and Brain Sciences* 3: 417-424.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind.* (Cambridge, Cambridge University Press).

Searle, J. (1992). *The Rediscovery of the Mind.* (Cambridge, Mass., MIT Press/Bradford).

Sejnowski, T. (1986). 'Open Questions about Computation in Cerebral Cortex.' In McClelland and Rumelhart (1986).

Sejnowski, T. and Rosenberg, C. (1987). 'Parallel Networks that Learn to Pronounce English Text.' *Complex Systems* 1: 145-168.

Shallice, T. (1988). 'Information-processing Models of Consciousness: Possibilities and Problems.' In A. Marcel and E. Bisiach (eds.), *Consciousness in Contemporary Science.* (Oxford, Clarendon Press).

Smolensky, P. (1987). 'The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn.' *Southern Journal of Philosophy* 26 (Supplement): 137-161.

Smolensky, P. (1988). 'On the Proper Treatment of Connectionism.' *Behavioral and Brain Sciences* 11: 1-23.

Shepard, R. and Chipman, S. (1970). 'Second-order Isomorphism of Internal Representations: Shapes of States.' *Cognitive Psychology* 1: 1-17.

Smart, J. (1975). 'On Some Criticisms of a Physicalistic Theory of Colors.' In C. Cheng (ed.), *Philosophical Aspects of the Mind-Body Problem.* (Honolulu, University of Hawaii Press).

Smythies, J. (1994). 'Requiem for the Identity Theory.' *Inquiry* 37: 311-329.

Sterelny, K. (1990). *The Representational Theory of Mind.* (Oxford, Blackwell).

Swoyer, C. (1991). 'Structural Representation and Surrogate Reasoning.' *Synthese* 87: 449-508.

Tienson, J. (1987). 'An Introduction to Connectionism.' *Southern Journal of Philosophy* 26 (Supplement): 1-16.

Tye, M. (1992). 'Visual Qualia and Visual Content.' In M. Tye (ed.), *The Contents of Experience*. (Cambridge, Cambridge University Press).

Tye, M. (1994). 'Qualia, Content, and the Inverted Spectrum.' *Nous*

Tye, M. (1995). 'A Representational Theory of Pains and Their Phenomenal Character.' *Philosophical Perspectives 9: AI, Connectionism, and Philosophical Psychology*.

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"? A Survey of the Concept of Distribution.' In W. Ramsey, S. Stich, and D. Rumelhart (eds.), *Philosophy and Connectionist Theory* (Hillsdale, Lawrence Erlbaum).

Van Gelder, T. (1995). 'What Might Cognition Be, If Not Computation? *Journal of Philosophy* XCI: 345-381.

Von Eckardt, B. (1993). *What is Cognitive Science?* (Cambridge, Mass., MIT Press/Bradford).

Zeki, S. (1985). 'Colour Pathways and Hierarchies in the Cerebral Cortex.' In D. Ottoson and S. Zeki (eds.), *Central and Peripheral Mechanisms of Colour Vision*. 19-44. (London, MacMillan).

Zeki, S. (1993). *A Vision of the Brain*. (Oxford, Blackwell).