real knowledge

the problem of content in neural epistemics

jan sleutels
Real knowledge
The problem of content in neural epistemics

een wetenschappelijke proeve op het gebied van de Wijsbegeerte

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
woensdag 14 september 1994,
des namiddags te 1.30 uur precies,
door

Johannes Jacobus Martinus Sleutels

geboren op 20 januari 1960 te Meijel
promotor: prof.dr. A.A. Derksen

ISBN 90–9007373–6

© 1994 by J.J.M. Sleutels
Kronenburgersingel 265A, 6511 AS Nijmegen

All rights reserved. Except for the quotation of short passages for the purposes of criticism and review, no part of the material protected by this copyright notice may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, recording or otherwise, without written permission of the copyright owner.
Contents

Preface 5

Chapter 1
Beyond the naturalistic fallacy 7
1. The unreality of knowledge 7
2. Masters of the universe 9
3. Neural epistemics and mental content 12

Chapter 2
Philosophy of mind. An overview 14
1. Cognition and psychology 15
2. The functionalist framework 18
3. The computational paradigm 21
4. Troubles with functionalism 26
5. The modularity of mind 33

Chapter 3
Eliminative materialism and folk psychology 45
1. Eliminative materialism 46
2. Conceptual or metaphysical arguments 52
3. Methodological arguments 62
4. Empirical arguments 68
5. Folk psychology explained 74

Chapter 4
Neural epistemics 79
1. Elements of connectionist architecture 81
2. A conceptual universe for connectionist theory 92
3. Neural epistemics. The cognitive allure of connectionism 102
4. Connectionism and mental content 113
## Contents

### Chapter 5

<table>
<thead>
<tr>
<th>Qualia and epistemic content</th>
<th>122</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Qualia disqualified?</td>
<td>123</td>
</tr>
<tr>
<td>2. Weak qualia examined</td>
<td>131</td>
</tr>
<tr>
<td>3. Color qualia and the labeling fallacy</td>
<td>137</td>
</tr>
<tr>
<td>4. Strong plasticity of perception</td>
<td>142</td>
</tr>
<tr>
<td>5. Theory and observation</td>
<td>149</td>
</tr>
</tbody>
</table>

### Chapter 6

<table>
<thead>
<tr>
<th>Computationalism and narrow content</th>
<th>157</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Punch cards and mental content</td>
<td>158</td>
</tr>
<tr>
<td>2. Nine reasons for internalism</td>
<td>171</td>
</tr>
<tr>
<td>3. A Cartesian heritage</td>
<td>195</td>
</tr>
</tbody>
</table>

### Chapter 7

<table>
<thead>
<tr>
<th>Externalism and mental representation</th>
<th>201</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A general argument for externalism</td>
<td>202</td>
</tr>
<tr>
<td>2. Causal theories of representation</td>
<td>208</td>
</tr>
<tr>
<td>3. The taming of the shrew</td>
<td>214</td>
</tr>
<tr>
<td>4. Second thoughts about externalism?</td>
<td>231</td>
</tr>
</tbody>
</table>

### Chapter 8

<table>
<thead>
<tr>
<th>Real knowledge in perspective</th>
<th>238</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A short history of knowledge</td>
<td>239</td>
</tr>
<tr>
<td>2. Transcendental philosophy reconsidered</td>
<td>251</td>
</tr>
<tr>
<td>3. The bone of content</td>
<td>259</td>
</tr>
<tr>
<td>4. Back to the future</td>
<td>266</td>
</tr>
</tbody>
</table>

### Notes

271

### Bibliography

289

### Index

301

### Summary in Dutch

315

### Curriculum vitae

319
Preface

Research for this study was begun in 1987, when I started out as a postgraduate student of Ton Derksen at Nijmegen University, Department of Philosophy. Work on it was continued as I transferred to the Department of Philosophy at Leiden University in 1989, where I have since then been teaching metaphysics and theory of knowledge. Support from these institutions is gratefully acknowledged.

Some of the material used in this study draws on previously published papers of mine. Earlier versions of chapters 3 and 5 were published in Dutch as Sleutels 1988 and 1993, respectively. These papers were rewritten for the present study, including new material and adding new parts. Some of the ideas developed in chapter 4 draw on a joint paper with Bart Geurts, then of the IBM Research Centre at Stuttgart, Germany, which was published as Sleutels and Geurts 1989. Parts of chapter 2 were originally prepared for the *Handbuch Sprachphilosophie*, eds. M. Dascal et al. (Berlin, De Gruyter), which is still forthcoming.

Many teachers, colleagues, students and friends have helped to make this a different book from what it was when I first thought of it. I thank them for their inspiration and criticism as well as for manifold distractions.
Contents
Chapter one

Beyond the naturalistic fallacy

1. The unreality of knowledge

This book springs from a single worry: that our knowledge of the world may not be real, in the sense that it may not be part of natural reality. Philosophers inform us that knowledge is logically prior to the world grasped by it, and that, therefore, it is not possible to study knowledge from a natural point of view. I want to explore this impossibility. I believe that knowledge is real, and that it is imperative to study it from that point of view. By the same token, the subjects of knowledge are part of natural reality, not something alien and sublime. We human beings use our brains for thinking. If we want to study real knowledge, we must study brains.

A landmark in contemporary discussions of knowledge, and an important source of the ‘anti-realism’ mooted here, is Wilfrid Sellars’ criticism of the Myth of the Given (Sellars 1963). Sellars energetically opposed the idea that there is a privileged stratum of fact, facts that are taken as ‘given’, on which our empirical knowledge of the world is founded, and to which all meaning and truth can be reduced. Such ‘givens’, whether they are thought of as sense data or material objects, first principles, universals or propositions, are often supposed to act as the ‘unmoved movers’ of empirical knowledge. According to Sellars, all attempts to reduce knowledge to a privileged stratum of natural fact are doomed to fail. To make knowledge real in this sense is to commit a species of naturalistic fallacy:

“the idea that epistemic facts can be analysed without remainder—even ‘in principle’—into non-epistemic facts, whether phenomenal or behavioural, public or private, with no matter how lavish a sprinkling of subjunctives and hypotheticals is (…) a radical mistake—a mistake of a piece with the so-called ‘naturalistic fallacy’ in ethics” (Sellars 1963, 131).

The backbone of Sellars’ argument is his claim that knowledge is radically linguistic, while the alleged ‘givens’ are supposed to be not. According to
Sellars, knowledge is not a matter of being in a certain *empirical* state, such as being affected by an object in certain ways, or being in a certain brain state. Rather, it is being in a *logical* state, a matter of relating to one’s linguistic peers, of logically justifying one’s assertions in the face of the language community. Sellars sharply distinguishes these two aspects of knowledge attribution:

“in characterizing an episode or state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says” (*op. cit.*, 169).

Succinctly put, if I believe that apples are red, this fact cannot be analyzed into intrinsic facts about apples, nor into sensory impressions, behavioral dispositions, or intrinsic features of my brain. My belief counts as *knowledge* just because my linguistic peers are typically prepared to accept my claim that apples are red. Knowledge is located in the ‘logical space’ of language; any attempt to analyze it as a natural process falls prey to the naturalistic fallacy.

*Consequences of the linguistic turn*

Sellars’ criticism of the Myth of the Given exemplifies the linguistic turn in modern philosophy, other exponents of which are W.V. Quine (1953; 1960) and Ludwig Wittgenstein (1953). More recently, the linguistic turn was forcefully epitomized by Richard Rorty in his influential book, *Philosophy and the mirror of nature* (1979). Rorty drives home the point that knowledge is not “a transaction between ‘the knowing subject’ and ‘reality’” (Rorty 1979, 9). It is not something real, but something located in the self-contained sphere of discourse. According to what may be called the ‘linguistic consensus’ in contemporary philosophy, all efforts “to break out of discourse to an archè beyond discourse” are fruitless (Sellars 1963, 196; cf. Pols 1992, ch. 3).

The linguistic turn holds some remarkable consequences for our conception of man, world, and knowledge. First, it invites us to *sublimate* the subject of knowledge as being no longer the empirical human self, but an abstract node in the web of discourse. The human self tends to become a mere “contributor” to the “ongoing conversation of mankind” (Rorty 1979), or a “center of narrative gravity” (Dennett 1992). The self is thereby exiled from the world known by it. It is a serious problem in the philosophy of mind how to account for this unworldly nature of the subject.

Secondly, the linguistic turn fosters a tendency to *evaporate* reality as an
independent constraint on knowledge. In the final analysis, the world’s contribution to knowledge is reduced to a series of ineffable surface irritations. The world can no longer be claimed to determine either the meaning, truth or validity of our empirical statements. This poses a severe threat to empiricism. In Rorty’s well-known phrase, the world now seems to be well lost (Rorty 1972).

Finally, the linguistic turn suggests that knowledge as such is *not* real. Knowledge becomes a thing alien and sublime, something not-of-this-world, as opposed to the world grasped by it. This is a serious problem for any aspiring cognitive science. The professed aim of cognitive science is to study knowledge as an empirical phenomenon. But if Sellars’ criticism of the Myth of the Given is correct, no empirical account of knowledge can succeed, because knowledge is *not* an empirical reality.

I am deeply troubled by the Myth of the Given, but I am no less troubled by the above consequences. The idealist tendencies released by Sellars’ argument, although appealing from a humanistic and Romantic point of view, strike me as fundamentally unacceptable. In this study, I want to argue for a more balanced conception of knowledge, such that *both* the reality of knowledge and its logical or epistemic aspects can be discounted. The problem, then, is how to study cognition as an empirical phenomenon without giving way to the Myth of the Given.¹

2. Masters of the universe

The issue raised here harks back to an age-old battle of disciplines. It reminds of the controversy between metaphysics and epistemology fighting for the position of ‘first philosophy’. Who should be Master of the universe? Should a theory of knowledge form part of the overall theory of reality (metaphysics), or is there a special discipline (epistemology), logically prior to all accounts of reality, which studies knowledge as a realm *sui generis*? As is well-known, Aristotelians in Antiquity and in the Middle Ages generally opted for metaphysics as first philosophy, while Descartes and Kant transferred this title to epistemology.²

Current approaches in cognitive science continue the battle of philosophies. Broadly speaking, there are two competing research programs. The first pays tribute to the rationalist tradition of epistemology, emphasizing the abstract and linguistic nature of cognition. It is represented most prominently by thinkers such as Jerry Fodor and Zenon Pylyshyn. On this approach, cognition is located in the ‘logical space’ of computations, which are viewed as purely
abstract functions of their physical substrates. The second program follows the naturalistic line of metaphysics, emphasizing the need to study mind and cognition as parts of natural reality. Boldly put, this line claims that cognitive science is the science of the brain. Advocates have even suggested that we should eliminate the traditional conception of knowledge, and supplant it with neurophysiology. The naturalistic program is represented most prominently by thinkers such as Paul Churchland and Patricia Churchland.

A Cartesian presumption

Both of these programs fall prey to the Myth of the Given, although for different reasons and in different respects. They both try to secure, in Sellars’ words, “a privileged stratum of fact” in terms of which to account for cognitive phenomena. Fodorians have mental symbols, naturalists have brains. Both symbols and brain states figure as ‘given’ in these theories, that is, they serve as the intrinsic basis of cognition. The main difference between the two approaches is that symbols are primarily logical items, whereas brains are physically real. If, as I have intimated, cognitive science should take into account both the logical and the real aspect of knowledge, then clearly a meeting of minds is in order.

This study explores the possibility of combining the insights of computational and naturalistic approaches to cognition. In this respect it contains nothing new, apart from certain accents in the manner of presentation. Similar proposals for a neurocomputational approach have been made by others in the recent past (see, for example, Churchland 1989a; Clark 1989; Churchland and Sejnowski 1992). What is new, however, is my attempt to combine the approaches in a coherent way that enables us to move beyond the naturalistic fallacy identified by Sellars. So long as cognitive science remains wed to the Myth of the Given, it will never be able to free itself from charges of reductionism and foundationalism. This becomes particularly clear in discussions of mental content, where it is typically argued that the semantics of cognitive states must be determined by intrinsic factors if these states are to count as cognitive. This presumption appears to be shared by rationalists and naturalists alike. Whether the content of belief is thought to be fixed by the combinatorial syntax of mental symbols, by synaptic arrangements in neural nets, or by causal or counterfactual relations to properties of distal objects, it is destined to be treated as given, that is, as being dictated to us by one of the Masters of the universe. A truly balanced account of knowledge, or so I shall argue, requires that we give up this idea.
Beyond the naturalistic fallacy

Seen from a slightly different perspective, the present study is a critique of the representationalist framework of cognitive science. Both the symbolist program of cognitive psychology and the connectionist or subsymbolist program in neuroscience are committed to the idea that cognition is the processing of internal representations (cf. Cummins 1989). Now, a closer look at the notion of representation as it is used in these theories reveals a fundamental reductionist presumption, namely, that representations must be given in order to count as cognitively relevant. In the terminology developed in chapter five, there is a tendency in cognitive science to make representations either ‘superstrong’ (endowed with intrinsic content) or ‘superweak’ (cognitively irrelevant). For example, a familiar objection against computational models is that computations as such are meaningless: having no content of their own, they must be irrelevant for our understanding of knowledge. Similarly, neural models of the mind are often charged with ‘changing the subject’. Neural activity has no intrinsic meaning; therefore, neuroscience cannot explain cognition, although it may have many things to say about the anatomy and physiology of the brain. I think that the presumption of intrinsic content underlying these arguments is wrong. It belongs to an essentially Cartesian frame of mind that is deeply entrenched in modern philosophy. Descartes inculcated us with the notion of man as an outside observer, a thinking substance locked inside a biological machine. The Cartesian subject has access to the external world only through his mental representations of that world. These representations are given to the internal subject, but they are not physically real. I shall argue for the exact opposite of this view, namely, that representations are physically real and that they are not given.\(^3\)

Notice that my concern with ‘givenness’ and with the naturalistic fallacy, unlike Sellars’ and Rorty’s, is not based on linguistic or social considerations. Social constraints on knowledge appear to me to be very much on a par with other external factors weighing on internal structure. I take them to be real and important, but not special. Habits, practices and institutions, most notably including those of language communities, control the formation of structured responses and behavioral dispositions in individual human beings. Language is a behavioral disposition like others, a set of neural structures specifically calibrated to respond to certain external pressures. Hence, an adequate understanding of language requires a more general theory of how physical and social mechanisms alike bear on the individual’s performance. From this purview, it would be methodologically unwise to start with language, or to think that only language matters.
3. Neural epistemics and mental content

My purpose in this book is to show that a theory of neural epistemics, based on modern connectionism, is both possible and desirable. To achieve this end, I examine various aspects of cognition and mental content, including the status of folk psychology, the power of connectionism, and the vicissitudes of content in a computational and physicalist context. A substantial portion of the book is devoted to what is nowadays called ‘psychosemantics’ (Fodor 1987) or ‘biosemantics’ (Millikan 1989): the theory of meaning for mental symbols required by cognitive science. My contention is that a global reorientation with regard to content is needed if we are to overcome traditional objections to naturalized theories of knowledge. I argue in some detail how this reorientation can be brought about.

A survey of the argument

The plan of the book is as follows. In chapter two, I offer a brief introduction to some of the main issues in contemporary philosophy of mind. I rehearse the development of reductionism, behaviorism, and functionalism in philosophy and in psychology, and review the main problems and strengths of these positions. I then turn to what is now the received view in philosophy of mind, computationalism-cum-functionalism. Some important aspects of this view are discussed, including the intentionality debate instigated by Searle’s Chinese Room argument, and the issue of modularity raised by Fodor and others. Together, these subjects provide the necessary background for the discussion in subsequent chapters. Readers already familiar with the terrain may want to skip this part. I advise them to read only the last section of chapter two, which contains some programmatic remarks on the relationship between science, metaphysics, and naturalism.

Chapter three examines various aspects of Churchland’s theory of eliminative materialism—metaphysical, methodological, as well as empirical. I explain what it is that we should forswear if eliminative materialism is correct, and why we cannot afford to do so. A ‘dual aspect’ theory of the propositional attitudes is introduced to account for the leading role played by folk psychology in our conception of cognition. On this basis, I want to make a case for large-scale integration and conceptual interaction between top-down research and bottom-up constraints.

The discussion of conceptual interaction is continued in chapters four and five, where I consider in some detail the explanatory resources of connection-
ism. I try to show that connectionism is able in principle to integrate our scientific understanding of the brain with folk psychology’s descriptive vocabulary. An examination of the main objections raised against this claim by connectionists as well as by their opponents reveals that they rest on a misconception of folk psychology and of the nature of mental content. I identify this faulty doctrine as a form of closet Cartesianism: the presumption that epistemic content must be determined intrinsically. In chapter five, I focus on a concrete example of mental content that is traditionally seen as particularly resistant to physicalist explanation, namely, qualia. I present a modified and expanded version of Churchland’s connectionist account of qualia, and show that it can deal with traditional objections on the condition that the presumption of intrinsic content is abandoned.

The doctrine of intrinsic content is scrutinized further in chapters six and seven. I trace its roots in traditional forms of physicalism and computationalism, and pursue its conceptual ramifications in causal theories of representation. Two competing research programs are distinguished, internalism and externalism. I argue that intrinsic content is a central tenet of internalism, but that the arguments adduced in its behalf are seriously flawed. Although externalism holds the better cards here, it ultimately succumbs to a form of intrinsic content, too, namely, by tending to construe mental content as being remotely controlled by the environment. I press the need for a more balanced account of content and cognition, such that the contributions of both the organism and its environment can be discounted.

In the final chapter, I discuss the issue of content from a broader historical and philosophical perspective. Taking my lead from a generalized notion of transcendental deduction, I suggest a radical alternative to internalism and externalism, called ‘relationism’. The true significance of relationism lies in its rejection of all intrinsic content. By this token, or so I shall argue, it enables cognitive science to finally move beyond Sellars’ Myth of the Given, and toward a naturalism without the naturalistic fallacy.
Chapter two
Philosophy of mind. An overview

*The Balnibarbian knowledge machine*

On one of his amazing travels, Gulliver finds himself on the isle of Balnibarbi. There he is shown a curious contrivance that is claimed to enable “the most ignorant person at a reasonable charge, and with a little bodily labour, to write books in philosophy, poetry, politics, law, mathematics and theology, without the least assistance from genius or study” (Jonathan Swift, *Gulliver’s travels*, III, 5). It is a contraption of twenty feet square, the surface of which consists of a large number of wooden dice, linked together by slender wires and covered on every side with “all the words of the Balnibarbian language in their several moods, tenses, and declensions, but without any order”. Furthermore, the “strictest computation” has been made “of the general proportion there is in books between the numbers of particles, nouns, and verbs, and other parts of speech”. Gulliver reports that, at the operating professor’s command, forty pupils

“took each of them hold of an iron handle, whereof there were forty fixed round the edges of the frame, and giving them a sudden turn, the whole disposition of the words was entirely changed. He then commanded six and thirty of the lads to read the several lines softly as they appeared upon the frame; and where they found three or four words together that might make part of a sentence, they dictated to the four remaining boys who were scribes. This work was repeated three or four times, and at every turn the engine was so contrived, that the words shifted into new places, as the square bits of wood moved upside down” (*ibid.*).

This procedure is repeated time and again, and has already produced several volumes of broken sentences, which the professor “intended to piece together, and out of those rich materials to give the world a complete body of all arts and sciences…”

The author left little to be guessed as to what he thought of the project. Little did Swift know that, one day, his story would become true. In modern
cognitive science, the Balnibarian knowledge frame, forty boys and professor included, has a place inside the human head. While cognitive science does not aim to “piece together the arts and sciences” itself, its purpose is to understand how knowledge is produced by certain organisms and machines, conceived as devices very much like the Balnibarian knowledge machine.

The Balnibarian project relies on a number of highly specific conceptual tools brought together from a wide range of disciplines, including traditional epistemology, metaphysics, psychology, artificial intelligence, linguistics, computer science, neuroscience, and anthropology (for a cursory survey of the sources, see Gardner 1985). The conceptual foundations of computationalism are controverted issues in the philosophy of mind. In this chapter I discuss some important questions that have been raised in this context over the past fifty years. Thus, what kind of ontology of mind is required for positing a knowledge machine inside the head? Is this machine cut from the same biological cloth as the other ingredients of organisms? Is the knowledge machine the source of intentionality? Who or what is in charge of the machine, and how does he manage? Is the mind-machine really crunching sentences, as in Gulliver’s example, or is it rather calculating neural activity? My aim in this chapter is to give a first and fairly global overview of modern philosophy of mind. I want to introduce some of the basic concepts and positions, which may serve as the frame of reference for more detailed discussions in subsequent chapters.

1. Cognition and psychology

Modern linguistic theory is so complicated that it takes years of study to master. Yet, in principle, we have been familiar with the rules of grammar ever since childhood. There is a paradox here, in the case of linguistic knowledge as well as in that of other kinds of cognition. We have mastered more or less effortlessly the rules of reasoning, and we have all learned to use our eyes; yet, little is known about just how the mind does these things. Viewed from closer quarters, the processing of visual information is an incredibly complex task; it takes years of study to even begin to understand the neural and cognitive processes involved. As Richard Gregory put it, there is more to seeing than meets the eye. This remark generalizes to other feats of cognition: there is more to knowledge involved than commonly acknowledged.

Linguistic theory provides a formalism in which to describe the languages we use. This does not automatically mean that linguistic knowledge, as
it is presumably stored in the mind, corresponds to the rules of grammar as discovered by linguists. There is a bewildering proliferation of competing grammatical theories, from which it is difficult to choose. In addition to Chomsky’s ground-breaking theory of *Syntactic Structures* (1957), linguists have suggested different varieties of ‘standard theory’, government and binding theory, generative semantics, Montague grammar, so-called realistic grammar, arc-pair grammar and abstract phrase structure grammar, to name only a few. Which of these theories, if any, is correct, is an astutely debated issue.

It seems only natural to turn to psychology for an answer to this question. Arguably, the psychology of natural language processing may help to determine which of these grammars is the *real* one, that is to say, which one is actually employed by the human mind. Thus Chomsky has repeatedly claimed that we should “try to develop the study of linguistic structure as a chapter of human psychology” (Chomsky 1972, 66). *Pace* Chomsky, however, linguists as a rule do not seem to be overly impressed by psychological considerations; the marriage between psychology and linguistics, as McCauley (1987) has put it, is not a very happy one. The grammarians’ aim is to produce simple, concisely formulated grammars. But, of course, the elegance with which a set of grammatical rules is stated has little bearing on mental mechanisms that are subject to processing constraints of an entirely different order. Integrating the linguist’s elegant rules into a real time language processor may require computational procedures that are much too complex for its finite resources, whereas a slightly less elegant set of rules with the same consequences for linguistic structure may be computationally more tractable.

Much the same is true of other cognitive abilities. In the case of arithmetic, for example, the rules of formal number theory constitute the most elegant ‘grammar’ of calculation. Yet, these rules are unlikely to be the ones that are actually employed by the mind. People presumably use a much more messy set of internally represented procedures, such as tables for multiplication and addition, carrying rules, and various other shortcuts. Again, in the case of visual perception, the rules of projective geometry are unlikely to be the ones that are actually employed by the ‘intelligent eye’. Instead, the ‘intelligent eye’ is bound draw on various cues, conjectures and other cognitive shortcuts that are more easily implemented (cf. Marr 1982; Fischler and Firschein 1987).

One possible solution to the problems envisaged here hinges on the familiar distinction between competence and performance (see, for example,
Johnson-Laird 1983, 167). In the case of language processing, for example, we may say that linguistic theory, or grammar, specifies competence, that is, the function to be computed by the language processor. As is well-known from automata theory, any function may be computed by indefinitely many possible procedures, however. It is the task of psycholinguistics, then, to determine which of these procedures are actually used in linguistic performance.

Some writers question the viability of the competence/performance distinction, in particular with regard to linguistic theory. They hold that no clear boundary can be drawn here, and that in fact the abstract concept of linguistic competence is redundant. The psychological study of actual language use (performance), so they claim, leads to an account of language itself without the need of an independent ‘normative’ discipline of linguistics (thus, for example, Clark and Haviland 1974).

Others take the opposite stance, disputing the relevance of psychological theory to our understanding of language. Thus Jerrold Katz (1981) has argued for a view of linguistics in which it is completely independent of psychology. On Katz’s view, natural languages are abstract Platonic objects, which not only can be studied independently of any theory of actual linguistic performance, but which are really independent of language users. This buys the linguist complete theoretical freedom, but only at a very high price, namely, by denying linguistic theory all empirical significance. Much the same holds for cognition in general. If knowledge is construed as a purely logical phenomenon, located in the realm of Platonic being, then it must be void of empirical reality. On this view, knowledge is simply not real in the empirical sense of the word; accordingly, a natural study of knowledge is a contradiction in terms.

The Platonist position is offensive to the intellectual palate of many modern scientists of mind, including that of the present author. Insulating our understanding of man, as a cognitive being, from that of the rest of nature, is now widely thought to be unacceptable both from a metaphysical and from a methodological point of view. In the defiant words of Jerry Fodor, “go ahead, be a Platonist if you like. But the action is all at the other end of town” (Katz 1985, 160). The present study shares with most of the naturalist literature on the subject a decidedly anti-Platonist outlook. Many writers, however, combine their naturalist convictions with a depreciation of so-called ‘transcendental’ philosophy in general. I shall be less rejective of transcendentalism. Details will have to wait until further development of my view in subsequent chapters, but the following can be said at this stage. I think that
knowledge is both logical, or epistemic, and real. In addition to the ‘ideal’ or ‘logical’ reality acknowledged by Platonists, knowledge is also an empirical reality, which can be studied from a natural point of view. To my mind, it is equally mistaken to think that cognition is purely non-empirical, as to think that it is purely empirical. A central purpose of this study is to develop a more balanced account of knowledge, which allows us to keep in mind both aspects at once.

2. The functionalist framework

Behaviorism and central-state materialism

It is a strange philosophy that endows man with a knowledge machine of the Balnibarbian type. To Swift, the idea was probably preposterous. To modern cognitive science, no idea is more dear. The philosophy of mind underlying the approach is generally called functionalism. It is best seen as a reaction against two earlier and less permissive theories of the mental, behaviorism and central state materialism (see, for example, Fodor 1981; Rorty 1982; Lycan 1990, 3ff; Searle 1992, ch. 2). The core idea of functionalism is the insight that psychological terms such as belief, desire, pain, memory, and meaning need not be understood as a kind of shorthand for either neurophysiological or behavioral descriptions.

Behaviorism, closely affiliated in linguistics with Leonard Bloomfield’s view of grammar as the theory of grunts and noises, denies the existence of an inner aspect of mental states, defining mental states without remainder in terms of publicly observable responses to publicly observable stimuli. Our everyday psychological vocabulary, according to behaviorism, is actually a set of shorthands for behavioral descriptions (the locus classicus here is Ryle 1949). If you share my feelings for Poulenc’s lyrical work, for example, then according to behaviorism that is for you and me to share a set of behavioral dispositions, including the tendency to listen to Poulenc’s music, to state your predilection when prompted, to buy an occasional EMI recording of his songs, and to muse over the beautiful poems by Apollinaire. In addition to this metaphysical doctrine of behaviorism, which is usually called ‘logical behaviorism’, there is also a milder variety called ‘methodological behaviorism’. The latter is more or less agnostic with regard to the nature and existence of mental states; its sole concern is with the proper methodology for studying mental phenomena. It rejects introspection as a reliable source of evidence with regard to the mental, and proposes instead to consider only correlations between controlled stimuli and publicly observable behavior
Behaviorism did not fare well. It soon became clear what it means to ignore the introspectively documented internal aspect of mental states and processes. In the case pictured above, for example, it may well be that you do not display any of the characteristic behavior, and that you like Poulenc nonetheless. A taciturn nature, lack of time and money may well conspire to assure that you do not publicly express your feelings for Poulenc’s work. On the behaviorist account, this would mean that you and I could not possibly be said to share a love for Poulenc. There is clearly something wrong here. Even more devastating for behaviorism was the insight that it is simply impossible to specify mental states in terms of a finite list of publicly observable behavioral dispositions, such that no overt or covert reference is made to ‘mentalistic’ items. The list of conditionals necessary and sufficient for analyzing *loves Poulenc’s songs*, for example, seems to be indefinitely long, inasmuch as there is no finite way of specifying in advance the countless ways in which the disposition could be realized. Much the same point was made by Chomsky against Skinner’s behavioristic conception of grammar. Chomsky argued that behaviorism implicitly assumes that the sentences of natural language can be produced and analyzed by a finite state machine, that is, by a device containing no working memory. The assumption is demonstrably false, however: in finite state devices, there is no effective procedure for producing sufficiently complex forms of linguistic behavior (Chomsky 1959).

The second theory, central state materialism (also called reductive materialism, or identity theory) is the most straightforward of the several materialist theories of mind. It claims that mental states just *are* physical states of the brain. To be more precise, it claims that each type of mental state or process is identical with some type of physical state or process within the central nervous system. The history of science attests to a score of similar reductions; thus, our common-sense notion of temperature has been neatly reduced to the scientific notion of mean molecular kinetic energy. If we are going to take serious the claim that man is part of the physical world, we should be prepared to look for a similar reduction of our common-sense notion of man and cognition.

Although there is doubtless a kernel of truth in the identity theory, its claims as stated are obviously too strong. In order for you and me to be in the same type of mental state, your brain’s pattern of activity would have to be exactly the same as mine. Yet, our everyday notion of mental states apparently does not involve anything like this neurophysiological constraint.
(as argued most notoriously by Kripke 1972). Nor does the brain itself: it is a well-attested fact that if certain brain structures are damaged, resulting in mental impairment (as, for example, after a stroke), the brain is often able to utilize alternative neural structures, eventually restoring the impaired mental function (a phenomenon called ‘plasticity’). A particularly dramatic example of the failure of straightforward mind/brain identity is the work of neuropsychologist Karl Lashley. Studying the effects of lesions in the cerebral cortex on intelligence and learning in rats and monkeys, Lashley became convinced that cortical structures are largely ‘equipotential’ as regards the acquisition and storage of behavioral functions. Although allowing for a weaker sort of localization, his findings supported the conclusion that no specific neural structure is responsible for the storage of specific data or memory traces. Thus, Lashley came to formulate his ‘Principle of Mass Action’: loss of specific memories depends on the extent rather than on the locus of cortical lesions (see, for example, Gardner 1985, 260ff). A third and final objection to central state materialism is the fact that this position entails that it is impossible for someone (or something) with damaged brains, or with abnormal brains, or with non-human brains, or even with no brains at all (say, computers and Martians), ever to be intelligent, or to have a mental life like ours. For obvious reasons, this objection has come to be known as the argument from ‘species chauvinism’ (Block 1978).

Functionalism and token-materialism

These and other difficulties with behaviorism and with mind-brain identity theories led to the suggestion that mental states should be identified neither with a person’s behavioral nor with his brain states, but rather with the brain’s functional states. Functionalism, crudely put, is a way of giving each side its due. It agrees with the behaviorist’s idea of psychological states as some kind of connection between the organism’s inputs and outputs, independent of the particular physical realization inside particular organisms or machines. But on the other hand it also agrees with the identity theorist that the mental is something internal, and that we should look at the interrelations of internal states in assigning psychological predicates to physical substrates.

According to functionalism, mental states and processes are defined by their functional role in causally mediating between stimuli, responses, and other mental states and processes. In essence, functionalism is ontologically neutral; it is as compatible with forms of dualism and idealism as it is with physicalism. Thus Hilary Putnam, one of the intellectual fathers of function-
alism, suggested that it is psychologically irrelevant what stuff we are made of, whether it be soul-stuff or Swiss cheese, protoplasm or silicon chips (Putnam 1975, 302). This so-called multiple realization feature, which holds good for silicon chips and Swiss cheese as well as for variations of brain structure from one person to another, is generally considered to be one of functionalism’s main assets.

Although functionalism as such is ontologically agnostic, it is combined most often with a generally physicalist outlook in the philosophy of mind. Rejecting the traditional forms of type identity, which hold that each kind (or type) of psychological event is to be identified with a type of physical event (for the identity theorist) or behavioral event (for the behaviorist), functionalism may be paired with the weaker physicalist claim that each particular instance (or token) of a mental type is identical with a token physical state. It is only universal type-type identities that are rejected, whereas token identity is perfectly compatible with functionalism. Typically, the rejection of type identity is taken to imply that cognitive science’s level of analysis is distinct, in the sense of being conceptually independent, from that of the various physical sciences, including neurophysiology, which study the functions’ substrates. Being the study of the substrate’s function, cognitive science has its own irreducible laws and its own abstract subject matter (see, for example, Fodor’s classical statement of this view in his paper on ‘Special sciences’, Fodor 1974). The rejection of bridge laws connecting the categorial systems describing mental events and physical events is also known as anomalous monism (Davidson 1970; Quine 1990).

3. The computational paradigm

Homuncular functionalism

Where does the Balnibarbian knowledge machine come in? Functionalism informs us that what is important about the mental is its functional organization, that is to say, its causal role in mediating between a certain input (information from the world outside, and certain given mental states) and a certain output (behavior, and new mental states). By this token, theories about the mental need neither restrict themselves to publicly observable stimuli and responses (as in behaviorism), nor to neurophysiologically established brain structures (as in reductive materialism), although both restrictions remain as marginal constraints (psychological theory should be relevant for explaining behavior, and should not introduce entities that are neurally impossible). Within these limits, however, contemporary psychol-
ogy is free to posit whatever entities it sees fit to explain our cognitive abilities. Exactly how these mental entities are pictured is irrelevant, so long as their functionally salient properties are duly observed.

In the study of language, for example, psycholinguistics effectively posits a machine inside the human head, a device of such functional properties as will account for our linguistic abilities. Typically, this machine is taken to contain a device for analyzing incoming information (pattern recognition, speech analysis), the output of which is fed into various other processors for analyzing the morphological, lexical, syntactical, and semantic properties of the input. Needless to say, these subdevices are all taken to be complexly interconnected, as well as connected to various memory systems containing the linguistic knowledge needed for executing these tasks.

The general idea of this functionally organized language machine is illustrated in Figure 2.1, adapted from a recent textbook on psycholinguistics by Wim Levelt (1989). The cognitive task to be explained is that of the psycho-

![Diagram of the language processor](image.png)

*Figure 2.1: Part of the language processor*

The diagram shows the global functional organization of the mechanism that is presumed to be responsible for producing and understanding spoken language. (Adapted from Levelt 1989.)
logical production and understanding of language. To this end, the main task is analyzed into a hierarchy of subtasks, represented by the various submachines in the diagram. Now, with regard to each of the subtasks, the question of how they are executed repeats itself. Functional analysis apparently entails that there is a staff of subpersonal demons or homunculi who are collectively responsible for executing the main task. Each homunculus executes a subroutine in the overall scheme of cognition, like that of the ‘conceptualizer’, the ‘formulator’, the ‘articulator’ and the ‘librarian’ in Levelt’s model.

The homuncular organization of the mind bears obvious resemblance to the Balnibarbian device for producing sentences. A graphic example may drive this point home. Figure 2.2, adapted from Lindsay and Norman’s introductory textbook on cognitive psychology (1977), shows part of the human language processor, which is represented as a kind of office containing a large blackboard, a choice of demon personnel, and various aids and resources. A demon scanner on top of the blackboard passes incoming information to pattern recognition demons, each a specialist in recognizing particular features of written letters. Other demons are involved in figuring out which word is being scanned. Again, some are trying to work out the syntactic structure of the (as yet unfinished) sentence; others make use of these results in eliminating possible candidates for the next word to be processed. In executing their specific functions, the demons may not only communicate with each other, but they may also draw on a store of (morphological, lexical, syntactic, semantic, etc.) knowledge in long term memory, delivered by messenger demons. The collective activity of these teams of demons is coordinated by a central supervisor, reminiscent of Gulliver’s professor in speculative learning.

As intimated by the demon caricature, the functionalist ploy of positing the existence of a knowledge machine is far from being unproblematic. It may be doubted whether this strategy will really explain our cognitive capacities. On the surface of it, the explanation proffered is just vacuous. It commits an obvious homunculus fallacy, inasmuch as the capacity to be explained is ascribed to an internal mechanism or subsystem. The fallacy is a species of the category mistake: it tries to locate, within some substructure, events and processes that actually belong to a higher level of description. In particular, the cognitive processes in which persons are involved are relegated to a shady realm of subpersonal entities.
Discharging homunculi: the computational paradigm

Whether or not positing demons is actually fallacious depends on the functions these demons are supposed to perform, and on the further analysis of these functions. If the subfunctions are mere duplicates of the function to be explained, then this is obviously fallacious. But if functions are analyzed into subfunctions, each accounting for a strategically smaller part or aspect of the mothering function, and if the subfunctions in their turn are analyzed...
into sub-subfunctions, such that, ultimately, they can be understood in terms of functional ‘atoms’ (logical circuitry, flip-flop switches, or something similar), then the manoeuvre seems relatively benign. As Daniel Dennett put it,

“homunculi are bogeymen only if they duplicate entire the talents they are rung in to explain. (...) If one can get a team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behavior of the whole, this is progress. A flow chart is typically the organizational chart of a committee of homunculi (investigators, librarians, accountants, executives); each box specifies a homunculus by prescribing a function without saying how it is to be accomplished (one says in effect: put a little man in there to do the job). If we then take a closer look at the individual boxes we see that the function of each is accomplished by subdividing it via another flow chart into still smaller, more stupid homunculi. Eventually this nesting of boxes within boxes lands you with homunculi so stupid (all they have to do is remember whether to say yes or no when asked) that they can be, as one says, ‘replaced by a machine.’ One discharges fancy homunculi from one’s scheme by organizing armies of such idiots to do the work” (1978b, 123-124; italics in original).

Discharging homunculi from the functionalist scheme is the task of computationalism. The computational paradigm in cognitive science is best seen as a tightening of the explanatory bite of functionalist ontology. Its purpose is to pursue the top-down analysis of the knowledge machine down to its most basic components, such that the execution of cognitive tasks is eventually understood in terms of computational operations on formally defined symbols. There is an important analogy with modern digital computers here. At user level, the operation of computers is usually understood in terms of global tasks like playing chess or guiding a welding robot’s arm. At programmer level, the execution of this global task is understood in terms of various subroutines, which, in turn, consist of sets of command structures. The program is typically written in Pascal, Lisp, Prolog, or some other relatively ‘high-level’ language; advanced programmers may use functionally more fine-grained instruction sets such as C. Still deeper down, the computer’s executing a given program may be understood in terms of machine language, that is, the set of instructions for performing very elementary operations on very elementary symbols, such as reading the contents of a given memory address and adding it to the contents of the accumulator. When program
execution is brought down to the orderly manipulation of mere zeros and ones, the final step to electronic hardware is easily made. Now, similarly, the execution of mental functions is expected to be analyzable in terms of the mind’s program, which similarly reduces to a ‘machine language’ of thought, and ultimately to neural ‘wetware’. In the field of artificial intelligence it is indeed explored how we can program a computer to simulate the cognitive performance of the human mind. Depending on one’s philosophical credentials, the suitably programmed computer may either be said to *simulate* the mind, or to even *emulate* the mind. On the first reading, which John Searle has dubbed “weak AI”, the program is a mere model for mental performance. The second reading, called “strong AI” by Searle, seems to be the received view in computationalism today: it entails that possessing the right program is having a mind (cf. Searle 1980; 1992, ch. 2).

Modern cognitive science rests on thoroughly computationalist foundations (see, for example, Pylyshyn 1984; Haugeland 1985; Block 1990). Without these foundations, the functionalist ontology of mind endorsed by cognitive science would lose much of its explanatory appeal. Dictating a research agenda, computationalism offers a promissory note on how, in principle, cognition should be explained. Of course, empirical research is needed to decide whether this promise is actually borne out.

4. Troubles with functionalism

*Surrealism in the philosophy of mind*

Paradoxically, functionalism is criticized for some of its merits. As we have seen, it was considered to be an asset of functionalism that its constraints on type-individuating mental states are less restrictive than those of the identity theory. However, it has been argued by Ned Block that every brand of functionalism will be either too ‘liberal’ or too ‘chauvinistic’, by either granting mentality to objects that do not have it, or by being inappropriately exclusive with distributing mentality (Block 1978). Either way, our spontaneous intuitions about what is to count as a real mental system will be violated.

In rough outline, the argument is as follows. Suppose we type-individuate psychological states by reference to input-output functions and to internal state transitions, as is proposed by functionalism. Thus, two systems would be in the same psychological state if there is a level of description on which both systems can be described as instantiating the same functional economy. But, obviously, this view is too liberal, because it counts too many
things as mental systems: not only the human mind/brain, but also a suitably organized telephone exchange, a collection of mice, cats, and interconnected mouse-traps, or even a giant ant-hill. Now, apparently, there are only two ways to escape from this quandary. First, one may try to avoid such counterexamples by adding further constraints on cognitive systems, such as the requirement that they be equipped with sensory systems, that they possess certain behavioral repertoires, or that they pass some other test. Alternatively, one may require putative psychological systems to resemble human thought processes at a more fine-grained level of analysis. Either way, Block argues, bogus psychological systems will be ruled out only at the price of ruling out real ones as well.

One rejoinder to this argument is that if the alleged counterexamples were given in more detail, they would either betray their own impossibility, or our intuitions about the mental would become clouded. Thus, it has been pointed out that Block’s line of reasoning is tied to a rather coarse-grained level of analysis, and that it would be much more difficult to devise a similar argument on the level of more fine-grained functional properties (see, for example, Clark 1989, 34ff). In terms of Dennett’s distinction between intentional stance, design stance, and physical stance, Block’s argument is tied to the intentional stance, whereas it would lose much of its plausibility if we were to try to expand it to the design stance, let alone to systems considered from the physical stance (cf. Dennett 1978a, 1987). Unfortunately, however, this reply is only an argument from ignorance; it is less than conclusive.

But is Block’s own argument conclusive? His style of reasoning has certainly caught on. In the past twenty years, much creative effort has been devoted to devising new and increasingly sophisticated examples pro and contra functionalism. (A highly amusing cross-section of this science fiction is Hofstadter and Dennett 1981.) Personally, however, I am inclined to suspect that the present proliferation of surrealism in the philosophy of mind makes one thing increasingly clear, namely, that for each apparently devastating argument against functionalism there is an equally convincing counterargument. This is obviously not the way progress will be made.

At this point, at least two different evaluations of the situation suggest themselves. First, it may be argued that Block’s type of argument does not so much show that functionalism is wrong, but rather that our intuitions are muddled. It may well be that we simply have no definite intuitions with regard to the nature of mental states; hence the present deadlock. To this we may add that the history of science attests to a score of similar situations. When Galileo claimed that the Earth revolves around its own axis, he was
ridiculed by pointing out that the consequences of his view are intuitively unacceptable: birds, clouds, and stones thrown in the air would be forced to move at dazzling speed as the surface of the Earth moves away under them. Similarly, the Twin paradoxes entailed by Einstein’s theory of relativity strike many of us as counterintuitive, even today. In cases like these, the conflict between intuition and theory is eventually resolved by giving way to theory. Perhaps the situation in cognitive science is similar: perhaps the functionalist frame of mind will in due time become a second nature to us, and the old intuitions will simply be forgotten.

A different way to assess the same difficulty is by analogy to Kant’s antinomy of pure reason (Kant 1781, A426ff, B454ff). Kant showed that it is as easy to demonstrate by relentless logic that space and time are bounded, as to demonstrate that they are not; similarly, that matter is composed of indivisible parts is as easy to prove as that it is infinitely divisible; similar considerations apply to the alleged proof of causal determinism and to the so-called ontological argument for the existence of God. According to Kant, the same fundamental mistake is made in all of these cases: transcendental features of our knowledge of the world are projected onto the world as such and mistaken for real beings. The proper Aufhebung of these problems is to acknowledge the transcendental nature of time, space, causality, and necessity. Similarly, in the case of cognitive science, the deadlock with regard to the functional nature of mind may be caused by our mistaking transcendental aspects for real aspects. Arguably, functionalism with regard to mental states is a properly transcendental feature of cognitive inquiry, which should not be confused with the claim that mind-as-such is really a functional entity. I shall have more to say about the transcendental nature of cognitive research in the final chapter.

*Entering the Chinese Room*

A second objection against functionalism, and more specifically against computationalism, concerns the phenomenon of intentionality. As we have seen, mental processes are defined by functionalism as algorithmic operations performed on internal symbols. These symbols have representational content: they are ‘about something’. In his seminal paper on ‘Minds, brains, and programs’, John Searle (1980) has argued that functionalism cannot be correct as a theory of the mental, because it will never be able to account for the aspect of intentionality. In rough outline, his argument is as follows.

Imagine a person, John, locked in a room and given a large batch of Chinese writing. Suppose John knows no Chinese and may not even be able to
discriminate Chinese writing from other kinds of squiggles. Now he is given a second batch of Chinese idiograms together with a volume of rules for collating the second batch with the first. The rules, which are in English, enable him to correlate one set of symbols with another. With a lot of practice John becomes highly accomplished at correlating any set of Chinese idiograms with the right (according to the rules) ‘answer’ in Chinese. To the native Chinese speakers outside, communicating with the room is now indistinguishable from communicating with real native speakers of Chinese. Although John executes all the right rules, he does not understand Chinese; *ex hypothesi*, all he does is to follow the procedures for manipulating formal, uninterpreted scribbles.¹

This thought experiment is intended to show that intentional cognitive phenomena, such as understanding language, can never be fully explained by a functional specification couched in terms of purely syntactic operations, that is, in terms of procedures for manipulating formal symbols. A system may perform impeccably in manipulating symbols in accordance with a given set of rules, and yet, for all that, it may be unaware of the meaning of the symbols, of what they are about. Formal symbol manipulations by themselves are meaningless; they are not even symbol manipulations, since the symbols do not symbolize anything *by themselves*. Whatever meaning they may have, they have in virtue of the intentionality of the subjects sending in the input and interpreting the output—in the above example: the native speakers of Chinese themselves.

Searle’s argument has released a true torrent of often surprisingly emotional replies (see the commentary ensuing Searle 1980, 424-450). Some writers energetically support Searle’s claim, anxious to avoid the looming implication of functionalism that the only difference between persons and thermostats is ‘a matter of degree’. Others, by contrast, vehemently resist Searle’s conclusions, charging him with mystification regarding the nature of mind, accusing him of antiscientism, or ridiculing him as just another philosopher gone astray. Still others try to undermine our intuition that there is no knowledge of Chinese involved; they urge, for example, that peripheral equipment should be added, such as a television camera to deliver the input and a motor system to deliver the output, and that the entire room should be placed inside a robot’s head. If these additional constraints are observed, *then* the system would be like a real Chinese—or so these critics claim. Many of these responses have been anticipated and answered in Searle’s original paper. To this day, the issue is still the subject of vigorous debate among philosophers and cognitive scientists.
According to Searle’s own view of intentionality (1983, 262ff; 1992, ch. 4), mental states are a biological phenomenon, “as real as lactation, photosynthesis, mitosis, or digestion.” According to this ‘biological naturalism’, as he calls it, the mental is a product of the brain, much like milk is the product of the lacteal gland. Intentionality is secreted by the brain in hitherto unknown ways, that will one day be unraveled by neurophysiology. Searle emphasizes that he does not deny that our mind is some kind of machine; on the contrary, it is his view “that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as the brain” (1980, 424, first emphasis Searle’s). Unfortunately, he does not make clear what is meant by ‘causal powers’ in this context. More precisely, he does not make clear in what his position differs from functionalism. Considering that the latter is expressly a study of the brain’s (or more generally: the machine’s) causal powers in mediating between certain inputs and certain outputs, Searle’s position would rather seem to be a kind of functionalism itself. To be sure, this brand of functionalism is significantly different from the version usually mooted in this context. In the first place, it seems to be tied to a much finer-grained level of analysis than that of the grossly characterized systems discussed earlier in this chapter. Its focus of interest are the ‘micro-functional’ and ‘micro-computational’ capacities of neurophysiological components of the brain, as opposed to the global i/o-profile instantiated by the brain as such (see also Clark 1989, esp. 34ff).

Secondly, Searle’s analysis starts from the fact that the human brain is capable of producing intentionality; only when the relevant causal powers of human beings have been identified does it make sense to inquire whether these same causal powers can also be instantiated by other entities. Traditional functionalism, by contrast, reverses the methodological order: it takes abstract, disembodied cognition as its primary target, and relegates all questions of instantiation to the proper study of substrates.

My own position in this book will be largely a form of brain-based microfunctionalism, in the above sense of the word. Like Searle, I am deeply suspicious of all attempts to construe the mind as a purely abstract and disembodied entity. My reasons for taking this position are twofold. In the first place, the abstract approach advocated by traditional functionalism seems to me to be at odds with the desire to establish cognitive science as an empirical study. The project of devising mere abstract models of cognition—flow charts, computer programs—is badly underdetermined by empirical data, and almost by definition so. One of the few empirical constraints more or less standardly admitted in abstract cognitive design is reaction time, which is used
as a (slightly unreliable) measure of computational complexity (Pylyshyn 1980; 1984, 120-130). If the set of relevant constraints is not drastically expanded (for example, by adopting some form of microfunctionalism), then cognitive modeling seems doomed to remain an empirically vacuous enterprise. I come back to this aspect of functionalism in chapter three. In the second place, the ‘disembodied’ approach rules out the possibility that the nature of mind, as well as its contents, depend on conditions of embodiment, including the interaction between the embodied system and its concrete, physical environment. I think this is bad methodology. The question of whether body and environment are relevant for mind and cognition should not be decided on \textit{a priori} grounds alone. I prefer to keep an open mind in this matter. Indeed, in chapters six and seven it will be argued that body and environment play a much more important role than is standardly allowed by functionalism. (For criticism of the notion of ‘disembodied intelligence’, see also Lakoff 1987).

\textit{Eliminative materialism and connectionism}

A third line of fundamental criticism against functionalism has been launched by Paul Churchland (1981; 1988a, 43 ff). His position, called \textit{eliminative materialism}, will be the subject of the next chapter, so I confine myself here to some brief remarks. Eliminative materialism agrees with functionalism in rejecting traditional type-type identity theories. Its reasons for doing so are altogether different, however. Eliminative materialism despairs of finding a one-to-one match-up between the concepts of psychology and those of neuroscience, not because the former are somehow abstract from their physical realizations (as in the multiple realization argument), but because they do not correspond to anything in reality. The mind, as we usually understand it, is simply a ghost: it does not exist. Hence, neuroscientific accounts of our inner lives cannot and should not be expected to endorse categories that match up smoothly with the categories of psychology. The ghost will eventually be eliminated, rather than reduced, by mature neuroscience.

The criticism here is directed more particularly against our common-sense understanding of the mind, or, as it has come to be called, our \textit{folk psychology}. In ordinary discourse, we describe, predict and explain the behavior of ourselves and others as that of intentional subjects processing discrete symbols according to determinate rules. The functionalist analysis of mental states in cognitive psychology takes this folk picture as its starting-point, functionally decomposing the states and entities globally identified as ‘memory’, belief that P’, and ‘desire that not-P’, for example, into their more
fine-grained components. What eliminative materialism resents most in this account is its inherent intentionality. While Searle charges functionalism with a lack of intentionality, eliminative materialism fears that it suffers from an overdose of intentionality. By the eliminativist’s standards, the alleged ‘intentionality’ of the mental is fundamentally at odds with our basic materialist conception of man. It is just as anomalous as was the ‘vital principle’ in nineteenth-century biology. The latter has been eliminated by organic chemistry, and so also the former should be eliminated by neuroscience. In short, intentionality is a mythical feature of a non-existing thing, the mind.

As baffling as these claims may seem, there is some evidence that central claims of folk psychology are indeed mistaken in important respects. In folk psychology, we tend to think of ourselves as intentional subjects processing discrete symbols according to determinate rules. If you see that it is raining and take your umbrella with you as you leave the house, your behavior is standardly explained in terms of various beliefs, desires, and rules. The explanation involves, among others, a perceptually caused belief that it is raining (P), a standing belief, presumably stored in long-term memory, that walking in the rain without an umbrella tends to get you wet (if P and not-U then Q), a desire not to get wet (not-Q), and a set of distinct rules, presumably also stored in memory, which enable you to infer that you should either wait until the rain stops (not-P), or take an umbrella with you (U). The basic model here is roughly that of a classical computer processing inputs in accordance with an explicit program and pre-stored data. However, at least one avenue of research currently being explored by cognitive science seems to imply that this view is inadequate as a model of the mind. The new paradigm, which seems particularly congenial to the eliminativist view, is that of connectionism, or parallel distributed processing as it is also called. It does not make use of explicit rules and procedures for manipulating explicit mental symbols. It does not think of memory, for example, as an ordered set of pigeon-holes, each containing a particular content (in the above example, ‘P’, or ‘if P and not-U then Q’) that can be systematically retrieved with the help of clear-cut procedures. Instead, like in a hologram, memory is supposed to be distributed over a network of (parallel and interconnected) processing units, each unit participating in the representation of many distinct contents. Again, in the case of language processing, the human language device is no longer seen as a centrally coordinated application of rules to symbols, as in the office model discussed above. Rather, the rules of linguistics are now seen as mere global effects of the collective activity of scattered units in a
network. The rules according to which these units operate and cooperate in no way resemble overt linguistic rules, nor do the ‘symbols’ processed by these units in any way resemble the overt (morphological, lexical, syntactic, etc.) symbols of traditional linguistics. (See, for example, Rumelhart and McClelland 1986a; Patricia Churchland 1986, 45ff.)

If connectionism is correct, the mind behaves merely as if it is rule-governed, as if it is crunching internal sentences, beliefs and desires, in a word: as if it is processing symbols. In reality, however, rules and symbols do not exist; therefore, they do not make up the mind. Mental activity is the product of many scattered units, each participating in numerous mental processes at the same time. These processing units are best thought of as idealized brain cells. As a matter of fact, much of the appeal of connectionism is due to the relative ease with which its models can be imagined to be models of the human brain.

Although some aspects of connectionism are obviously close to eliminativism (for example, its appreciation of neurological research, and its corresponding depreciation of folk psychology), this is not the only possible way to understand the new trend. Thus, Pinker and Prince (1988, 75-78) discern three possible relationships between connectionism and traditional symbol processing. One of them is straightforward eliminative connectionism. A second way to evaluate the bearing of connectionism on classical cognitive theory is to endorse implementational connectionism: on this view, connectionist models describe how traditional symbolic models may be implemented in a certain type of parallel hardware. Finally, there is a range of intermediate possibilities that may broadly be termed revisionist symbol processing cum connectionism. On this view, cognitive science may hold on to the idea that mental processes are essentially a matter of rule-governed symbol-processing, but not necessarily like anything we would spontaneously think of in terms of folk psychology. The rules need not resemble anything like the familiar rules of grammar, for example, and the symbols need not resemble letters, words or sentences; rather, these rules and symbols may prove to be complex wiring diagrams of networks of processing units.

A major aim of this study is to come to terms with the interrelations of connectionism, neuroscience, folk psychology, and classical cognitive science. As intimated earlier, I am dissatisfied with the view that knowledge is merely a logical or functionalist abstraction, a disembodied spectre void of all natural reality. But I think it is equally misguided to try to understand knowledge from an exclusively neurophysiological point of view, in the
way suggested by eliminative materialism. Developing a radical alternative to these positions will be an important task of chapters three and next, in which I argue for a theory of natural epistemics.

5. The modularity of mind

A final issue to be discussed here, and one that is again closely related to functionalism in the philosophy of mind, addresses the question of whether all cognitive processes are cut from the same cloth. If the mind is viewed as a functionally organized entity, it is still an open question how this organization bears on the proper architecture of mind. One possibility would be that the mind is indeed a seamless fabric, a monolithic whole characterized by a single, global input/output profile. Specific mental states are mere (sets of) instances of this global function. On this view, perceptions and beliefs, for example, are functions of the self-same global structure. Radically opposed to this unitarian view is the claim that the architecture of mind consists of distinct and highly specialized modules, soldered together, and communicating with each other in only very limited ways. While the unitarian position seems to be the more traditional one, and closest to folk psychology, the second or modular view is rapidly gaining in popularity among cognitive scientists. It has made its appearance under many guises in different fields. In computer science it is present in the concept of structured or modular programming. In psychology it appears as the notion of separate subsystems of cognitive architecture (for example, Marr 1982, 8-38). More specifically, in the field of linguistics it underlies Chomsky’s conception of a ‘language organ’ and of the ‘components of grammar’, that is, of a special device like the language frame described by Gulliver, made up, at a more fine-grained level of analysis, of relatively separate subsystems utilizing different types of linguistic knowledge (Chomsky 1980). In recent philosophy of cognitive science the idea of modularity has been elaborated in greater detail by Zenon Pylyshyn (1984, esp. 130ff) and Jerry Fodor (1983).

It is especially Fodor’s notion of ‘informationally encapsulated modules’ that has caught on. His theory of modularity is presented against the background of faculty psychology generally. Roughly, there are two kinds of faculty psychology, one horizontal and one vertical. According to the horizontal variety, most (if not all) cognitive processes make use of most (if not all) of the mind’s faculties, such as memory, attention, perception, sensibility, imagination, and so forth; the character of each such process is determined by the particular mix of faculties it draws on. The important point here is
that, on this view, it is the self-same faculty that is involved in the various
domains of cognition. For example, it would be the same faculties of judgement
and memory that are used in perceptual recognition (judging whether
a set of sensory data matches certain categories that are stored in memory)
and in deciding whether to accept a bid on your car. Horizontal faculty psy-
chology is presumably closest to the way we tend to think consciously of
how the mind works. But, as eliminative materialism has cautioned us, en-
trenchment in folk psychology is no guarantee for truth.

The vertical variety of faculty psychology, traced back by Fodor to Gall’s
work on phrenology, denies the existence of horizontal faculties shared by
all mental processes. It claims that there is no one such thing as a faculty of
memory, or of judgement, etc. Instead, there are many separate faculties,
one for each domain of cognitive processing. These faculties do not share re-
sources such as memory; the musical faculty, for example, has its own spe-
cific memory for music, which is quite independent of other types of me-

Now, Fodor proposes a view of mind that is of mixed horizontal and ver-
tical origin. He distinguishes between a large block of central processes and
a number of peripheral systems. The central processes make use of shared,
horizontal faculties; the peripheral systems are organized as vertical facul-
ties or ‘modules’, as Fodor calls them, each being computationally self-suffi-
cient and domain-specific. These modules include, roughly, certain compo-
nents of the input systems (most notably of visual perception and of lan-
guage understanding), and certain components of the output systems (in-
cluding motor control and language production).

Fodor discusses several features which he takes to be characteristic of
modular systems. Thus, modular systems are informationally encapsulated,
that is, they have access to only the information represented within the local
structures that subserve it, not to other background knowledge. Typical ex-
amples of this phenomenon are optical illusions such as the Müller-Lyer il-
lusion (see figure 2.3). Even when measurement has assured us that the two
arrows in the picture are of the same length, we still perceive one arrow as
longer than the other; background knowledge is not accessed by the visual
processing of the image. Secondly, the operation of modules is mandatory,
that is, modular systems perform automatically when appropriately stimu-
lated. We lack the ability to prevent modules from computing their func-
tions. In the case of language, for example, we simply cannot help hearing
an utterance of a sentence in our native language as a sentence rather than an
uninterpreted stream of sounds. In the third place, modular systems are do-
main specific, that is, they operate only on specific subclasses of stimuli. For example, there is some experimental evidence for the view that the computational systems involved in speech perception are triggered only by acoustic signals that are taken to be utterances. Finally, modular systems are hard-wired, that is, they are associated with specific neural structures. Consequently, modular mechanisms may be expected to display relatively specific breakdown patterns. Neuropsychological research seems to corroborate this view (see, for example, Luria 1973). Local brain lesions are associated with loss of mental function, especially in the case of perceptual systems and language processing mechanisms (agnosia, aphasia, etc.). Central processes, on the other hand, do not appear to be similarly localized; crudely put, there is no apparent neural structure for modus ponens.

As a result of these features, modular systems work very fast. Because their operation is mandatory, no deliberation is required to set the module into action. Because modules are domain specific, they may trade on fortuitous features of the domain without bothering about the validity of these features in other domains. Because they are encapsulated, the amount of information that needs to be taken into account is relatively limited. And because modular systems are hardwired, the requisite processing is drastically reduced.

![Figure 2.3: Optical illusions corroborate informational encapsulation](image)

Optical illusions corroborate informational encapsulation (see Fodor 1983, 47ff). From left to right: the Müller-Lyer illusion, the illusion of subjective contours, due to G. Kanizsa, and the Ponzo illusion. In the Ponzo illusion, apparent depth alters the perception of size. The perspective effect induced by the converging lines causes our visual system to make size corrections for the three-dimensional phenomenon of change in size with distance. In two-dimensional reality, the horizontal lines are of the same length. In the illusion of Müller-Lyer, the visual system makes similar corrections; in reality, the two vertical lines are of the same length. Illusions such as the above suggest that the processing of visual information is both mandatory and encapsulated from top-down contextual information. (See also Gregory 1981, 395ff; 1987, s.v. illusions; Fischler and Firschein 1987, 226ff).
Now, according to Fodor, these characteristics do not apply to central processes. Central processes somehow bring together the information from the various input systems; hence they cannot be domain specific. Moreover, they are not informationally encapsulated, since they somehow manage to combine information from a wide range of domains. In principle, each bit of incoming information may be brought to bear on the assessment of information from each of the other sources; discoveries in astrophysics may shed sudden light on problems in quantum physics or biochemistry. This holistic trait makes it very difficult to understand central processes. We are faced with what in Artificial Intelligence is called the ‘frame problem’, the problem of putting a ‘frame’ around the set of beliefs that need to be revised in the light of newly arrived information. If the processing of all data depends on all other data (Fodor speaks of the ‘Quinean’ or ‘isotropic’ character of central processes here), then central processes defy any principled analysis.

The modularity thesis has elicited much comment recently. In essence, it defines a research agenda, and it can therefore be criticized on empirical, methodological, as well as philosophical grounds. I give an example of each of these lines of criticism here. (For further discussion, see Gopnik and Gopnik 1986; Garfield 1987; Meijering 1989b; Ross 1990.)

One important field of empirical research on modularity in natural language processing is concerned with our ability to recognize the words in a text or utterance. A major issue here is whether lexical processing is contingent upon the linguistic and extra-linguistic context in which a word occurs. There are now two main theories, one modular, the other interactive. The modular theory develops the idea that lexical processing is an autonomous subsystem in language comprehension. It draws on the fact that in skilled language use word recognition processes are automatic, in the sense that they are very fast, occur without conscious effort, and do not seem to interfere with other processing tasks. Now, if these processes are in fact located within a modular subsystem, and are thus informationally independent of other aspects of language processing, two more predictions can be made about lexical processing. First, the information contained in the module’s output should be invariant across contexts. And second, the speed with which that information is made available should be unaffected by the processing context.

Both predictions are starkly denied by the interactive models of lexical processing, in which contextual information is combined with sensory information throughout lexical processing. According to this latter view, sensory processes are primed by top-down information about the lexical candidates
that would be congruous with the processing context, so that words in rich contexts will be processed more rapidly than words in impoverished contexts. Also, different information may become available in different contexts.

These contradictory predictions bear directly on the question of modularity. There is no conclusive empirical evidence on either side, so far. (For more details, see Seidenberg and Tanenhaus 1986. A conciliatory position between modular and interactive models is defended by Tanenhaus, Dell and Carlson 1987.)

A philosophical objection to the modularity thesis concerns the possibility that cognitive architecture is thoroughly isotropic and Quinean in character. According to Fodor’s ‘First Law of the Nonexistence of Cognitive Science’ (1983, 107), this would mean the end of cognitive science: we may be able to understand modules, but if there are no modules, it may well be that we will never be able to understand anything of how the mind works. This rejection of modularity, but without the ensuing pessimistic conclusion, is endorsed by John Anderson. An outspoken defender of a unitary view of cognitive architecture, he argues that there are no vertical faculties, but only horizontal ones (1983, xxff). Thus, he points out that many cognitive functions have only a very short evolutionary history, a history that would seem to be too short, in fact, to have produced specialized mental organs. Moreover, he argues that the faculties that are supposed to contribute to cognitive activities seem to be terribly intertwined. Most cognitive activities appear to have many features in common (like Fodor’s horizontally organized central systems), so that we would be hard put to trace their distinctive contributions.

Finally, I want to mention a line of methodological criticism. It is not inconceivable that the supposed modularity of a cognitive system is really nothing but the unintended product of adopting a certain research strategy. A large majority of the work in cognitive modelling is in fact concerned with small, specialized and relatively encapsulated parts of some hypothetical larger processing structure. Typically, this encompassing structure is represented only by way of some input/output profile and bookkeeping routines that feed and control the subsystem. This may look like a module in Fodor’s sense, but it need not be one; its encapsulation is probably only a by-product of the specialization of research.

Especially in the case of psycholinguistics the danger envisaged here would seem to be very real. Because linguistic theory is so extremely complex, there is a strong tendency to think of our capacity for language as a distinctive, specialized faculty. And there is a concomitant tendency among
psycholinguists to balkanize psychology by insulating theories of the language organ from other psychological research. Although excellent as a mere division of labour, however, this may well be a misrepresentation of the relationship between linguistic constructs and psychological processes generally, as is argued by, for example, William Marslen-Wilson and Lorraine Tyler (1987; see also Lakoff 1987; McCauley 1987).

The notion of modularity will return in particular in chapters four and five. One of the new ideas to be introduced there is that of ‘micromodularity’, a conceptual extension of modularity roughly analogous to the micro-functionalism introduced above.

6. Metaphysics of mind. A conceptual universe

Towards a naturalized metaphysics

The foundations of cognitive science are a chapter in the metaphysics of mind. I take the idea of a metaphysical inquiry in the sense of categorial analysis: its aim is to scrutinize the basic categories in terms of which we tend to conceptualize the world in a given domain of discourse, including science, common-sense and philosophy. This fairly straightforward notion of metaphysics is widely accepted in Anglo-Saxon philosophy, to the extent that it has textbook status (see, for example, Carr 1987; Hamlyn 1984). In this section, I want to highlight an important consequence of this view of metaphysics, and add some corrections to it. The consequence is the reflective nature of metaphysics, and the corrections concern the claim that metaphysics is merely reflective. The impact of my remarks will be a rapprochement of metaphysics and naturalism.

Metaphysics in the sense of categorial analysis is a reflective activity by definition; hence, it is essentially dependent on that upon which it reflects. A reflection on the categories deployed by our everyday discourse on persons, for example, or by evolutionary theory, depends on common sense and on scientific theory. In consequence of this dependency, metaphysics is required to be in continual contact with a wide range of conceptual activities. It is therefore considerably less other-worldly or weltfremd than is sometimes thought. On the contrary, if any one discipline cannot afford to be weltfremd, it is metaphysics. The metaphysicist is bound by profession to take good notice of any scientific and other developments that may affect his business.

These remarks are particularly apt in the context of the present study. As we have seen in this chapter, a pervasive feature of the issues discussed here is the amalgamation of philosophy, common-sense and empirical science.
Far from being situated in a cultural waste-land, modern philosophy of mind finds itself at the centre of a lively exchange of ideas from many directions. It is equally responsive to information from science and to ideas from common-sense, religion and traditional philosophy. The sciences that are most pertinent to the metaphysics of mind include psychology, biology and neurophysiology, of which several examples have been given.

The amalgamation of philosophy and science is usually defended in the name of naturalism; the above considerations suggest that it is equally defensible in the name of metaphysics. On the face of it, metaphysics differs from naturalism in at least one crucial respect, however. The notion of categorial analysis suggests that metaphysics is a merely reflective discipline; it stands severely aloof from that upon which it reflects, passively recording the inherent categorial structure of its subject-matter. Naturalism, by contrast, envisages an active participation of science in philosophy, and of philosophy in science. The metaphysicist is an outsider, the naturalist a participant. I think this view does not stand up to scrutiny, however, neither historically nor conceptually. It tends to underestimate the active role played by metaphysics in making categorial structure, as opposed to recording it. I want to suggest that we take the notion of metaphysics one step further toward naturalism.

The more traditional conception of categorial analysis hinges on a distinction between categories as ‘made’ and categories as ‘given’: science and its consorts may ‘make’ new categories to describe the world, but metaphysics is ordained to take these categories as ‘given’. Historically speaking, this is a very naive view of categorial structure. When Aristotle drew up a list of ten basic aspects of reality (which he called ‘categories’ or ‘predicaments’), when Descartes proposed his radical dualism of the physical and the mental substance with their concomitant hierarchies of attributes, or when Kant delineated the transcendental structure of sensibility and understanding, they were not merely explicating what was already there in the thought of their time. Rather, they helped to mould and change that thought. They did not simply give a name to the pre-existing categories of thought; rather, by naming them they in part created them. Their task was essentially one of bringing structure to the material on which they worked: sets of loosely related concepts needed to be ordered in a comprehensive system. The system as such did not previously exist, or existed only in a very incomplete form. Where clear relations between concepts were lacking, new relations needed to be established. A selection of relevant concepts was made, hierarchies between concepts were appointed, and questions were answered that had not been asked before.
Examples like the above strongly suggest that categorial structure is not ‘given’. Metaphysical analysis forges structure where no ‘ready-made’ structure existed before; conceptual schemes are ‘made’ rather than ‘found’. To be sure, categorial analysis is constrained by the set of concepts on which it sets to work; in this sense, it is certainly correct to say that the task of metaphysics is to enhance the structure latent in these concepts. The point, however, is that there are always indefinitely many of these ‘latent structures’ in any given set of concepts, each waiting to be ‘actualized’ by some form of categorial analysis. Yet, no single one of these structures can properly be claimed to be the ‘most correct’ or ‘real’ structure.

The aim of metaphysics, as explained here, is to enhance the conceptual self-awareness of a given domain of discourse. It does so not by explicating the unique conceptual scheme inherently used by that discourse, but by proposing a scheme, and by motivating this proposal. The choice of scheme will typically affect the manner of thought. The effect of categorial analysis will often be that our view of reality is revealed as confused and muddled in certain respects. Parts of the categorial structure appear to be simply not ‘filled in’, like blind spots in the visual field. In the case of scientific theories, categorial analysis may serve a useful purpose in tracing out a more robust conceptual scheme, in identifying weak points in current approaches, and in charting a possible course for further development and research. Metaphysics may inspire new research programs, and it may be used to discourage old ones as being conceptually stagnant or confused (cf. Lakatos 1978). In sum, metaphysical analysis falls in line with a genuinely naturalistic outlook on science and philosophy.  

---

A conceptual universe for cognitive science

To illustrate this point, let me give a concrete example of ‘naturalized’ metaphysics. As explained in this chapter, cognitive science and philosophy of mind present a wide variety of positions on a fair number of issues. An instructive way of visualizing this discourse is to think of it as carving out a ‘conceptual space’ analogous to the space of geometry. Specific theories are defined as vectors in this space, while clusters of theories, or research programs in Lakatos’ sense, may be represented as partitions or subspaces. Figure 2.4 shows a three-dimensional cross-section of the conceptual universe of cognitive science. I have singled out three degrees of freedom with regard to choice of theory: folk psychology, functionalism, and modularity. (For present purposes I assume these parameters to be conceptually independent; hence they are represented as orthogonal axes. Many other parameters
bear on the choice of theory; $n$ parameters define $n$-dimensional space, of which only a 3-D projection is shown here.)

In the bottom corners of the diagram four characteristic types of position are represented by four small volumes. In the left corner we find Churchland’s brand of eliminative materialism. Diametrically opposed to this position, and located in the right corner, is the implementational functionalism espoused, for example, by Fodor and Pylyshyn (notice that the parameter of modularity is left out for the moment; the reason for this will presently become clear). The various forms of reductive materialism (central state materialism, type-identity theory) are clustered together in the front corner, while the eliminative functionalism endorsed by Stich and Dennett finds itself in the rear corner of the bottom plane. Between these extremes a large number

---

**Figure 2.4: Part of the conceptual universe of cognitive science**

Only three basic dimensions of freedom in theory construction are shown, resulting in a three-dimensional projection of the conceptual universe of contemporary philosophy of mind. For convenience of exposition, it is assumed here that the dimensions are conceptually independent of one another; hence, they are represented as orthogonal axes. The theory of the modularity of mind may serve to bridge the gap between elimination and implementation of folk psychology and between mind-brain identity and classical functionalism. As explained in the text, the positions located in the four corners of the bottom plane have been drawn in without taking into account their stand (if any) on modularity; this affects in particular the positions of Fodor and Pylyshyn. Once the dimension of modularity is added, a new domain of arguments becomes available, making it possible to rationally combine moderate positions on the axes of functionalism and folk psychology.
of intermediate positions are possible. With regard to folk psychology, there is a graded spectre of revisionist positions between pure elimination and wholesale implementation. With regard to functionalism, the proper level of abstraction for cognitive science may arguably be anywhere between pure hardware and pure software. Finally, the parameter of modularity opens up a range of possible architectures for mental organization between the extreme positions of pure modularity and pure unity.

Figure 2.4 models categorial structure for contemporary philosophy of mind. Even though the model badly simplifies a complex situation, it is not implausible. Each of the positions mentioned in this chapter stands for a specific blend of categorial framework. Type-identity, for example, combines a specific version of functionalism with specific expectations about folk psychology. The model of figure 2.4 is typically the result of naturalized metaphysics in the sense outlined here. The analysis applied in it is not dictated by the theories it models. Consider the dimension of functionalism. The decision to situate reductive materialism and Fodorian cognitivism in the same dimension (that is, to identify them as forms of functionalism) is not the result of merely recording category structure; on the face of it, no two positions are more different than reductionism and cognitivism. The analysis calls for an act of creative interpretation, and goes beyond any ‘given’ structure. Taking the categories of substrate, function and causal i/o profile at face value, we would fail to see that they may be used both to defend the abstractness of mental states vis-à-vis their neural realization (the ‘multiple realization’ argument used in classical functionalism), and to defend a nearly reductionist, biological view of intentionality (the ‘causal powers of the brain’ argument in Searleian microfunctionalism). It is not so much the concepts that count, but the arguments and positions in which they are used.³

_A meeting of minds_

The structure of figure 2.4 is not forced upon us by the logic of the domain, although it may still be said to be constrained by it in a weaker sense. It is only one of many possible ways to structure the space of cognitive science.⁴ In the case at hand, my choice of parameters is dictated largely by purposes of exposition. This becomes clear as we turn to the dimension of modularity, which has been neglected so far. As outlined in the previous section, the question of modularity finds itself at the cross-roads of important questions in cognitive science. In figure 2.4 the orthogonal situation is taken in a literal sense. Modularity cross-categorizes the frameworks of folk psychology and functionalism, and is situated in a plane orthogonal to them. The added di-
mension opens up a range of new possibilities. A particularly interesting consequence is the fact that hybrid positions in the bottom plane can now be rationally justified. Seen from the two-dimensional perspective of folk psychology and functionalism alone, any particular combination of these two positions remains relatively arbitrary. Rationally speaking, hybrid solutions are less than satisfactory without principled reasons for adopting their specific blend of theory. The added dimension of modularity opens up a new field of concepts and arguments in terms of which to give such reasons.

If a meeting of minds is possible, it may as well be in a theory of the divided mind. Perhaps materialists and functionalists will meet halfway in the view that modular systems are amenable to reductive explanation, while the central system, and the global organization of the various modules, can be explained only in terms of abstract functions. A similar compromise may be found for defenders of elimination and of implementation. On the one hand, neurobiology may prove folk psychology wrong on module-related issues. On the other hand, following Fodor’s ‘First Law’, central systems may turn out to be strongholds of folk psychology, successfully defying all explanation in neuroscientific terms. Whether or not these compromising solutions are supported by the facts of cognition, only empirical research can tell. Metaphysical analysis informs us that they are conceptually sound and rationally defensible—hence, that it makes good sense to explore these fields of research.

The above suggestions are highly speculative, to be sure, and I will not pursue them any further here. Yet, they illustrate the course of ‘categorial counseling’ to be taken in the following chapters. If these are chapters in metaphysics, I hope to have demonstrated their potential significance for empirical science. If philosophers think they contain too much science, I hope to have demonstrated the latter’s metaphysical significance. By some quirk of history, philosophy is never allowed to harvest from the science it has sown. Whatever my merits in this respect, the fruits are no less real.
Chapter four

Neural epistemics

It has been noted that the term ‘scientific revolution’, since its formal introduction by Thomas Kuhn in 1962, has been used with increasing frequency in discussions of scientific change, and that the magnitude required for something to count as a revolution has diminished alarmingly (Ramsey, Stich and Garon 1991). This ironic comment seems to be particularly true of cognitive science, which in the course of mere decennia has seen at least two major revolutions—or so it is claimed. The first revolution was called the ‘Cognitive Revolution’ by chronicler Howard Gardner (1985). It was led off with a breakthrough in computer science, the development of the digital Von Neumann machine in the 40s and 50s. The second self-styled revolution is of more recent date. It was also initiated by developments in computer science, where in the late 1970s and early 1980s the possibilities of radically new architectures were explored, issuing in the construction of the Connection Machine (Hillis 1985). What emerged was a new paradigm, or more cautiously perhaps, a new research program in cognitive science, that is variously called ‘parallel distributed processing’ (Rumelhart and McClelland 1986a), ‘connectionism’ (Smolensky 1988), and ‘natural’ or ‘neural computation’ (Richards 1988; Churchland and Sejnowski 1992). I will be using these terms interchangeably here, though I will most often speak of ‘connectionism’.

Connectionism has seen an exponential growth over the past ten years, attracting the interest of scientists from a number of disciplines (Ramsey et al. 1991, ix). It appeals to neuroscientists as a model for real neural circuitry, while physicists pursue it for the useful analogies between connectionist nets and natural systems that display similar nonlinear dynamical behavior. Mathematicians study the formal descriptions involved, computer corporations are interested in commercial applications of brain-style processing devices, workers in artificial intelligence look to it for radically new ways of real intelligence, and psychologists find it attractive as a new account of human information processing. Finally, philosophers such as Paul Churchland and Andrew Clark turn to connectionism for an answer to the peren-
nial question of the relation between mind and brain.

From a sociological and psychological point of view, connectionism has all the ingredients for being a truly Kuhnian breakthrough. There is tumultuous interdisciplinary research activity, new journals and new societies are founded, while debates with opponents tend to end in confusion and mutual reproach. To a certain extent, the new approach is indeed *incommensurable* with the old ways of cognitive science: new ‘metaphors for thought’ are introduced (Norman 1986), new standards for evidential relevance and explanatory acceptability are put forward, new textbook paradigms are taught (Kuhn 1962, chs. 9ff). Thus, connectionists tend to concentrate on feature detection and pattern recognition, whereas cognitivists have been concerned primarily with language, reasoning, and other ‘higher’ cognitive functions. Apart from differences in the range of phenomena to be explained, the old approach and the new use different conceptual tools. The node-and-connection framework from which connectionists are working is quite unlike the familiar rule-and-symbol approach of cognitive psychology and GOFAI (‘Good Old-Fashioned Artificial Intelligence’; Haugeland 1985). To supporters this appears as a benefit of the new theory, while opponents see it a clear failure to address the proper explananda. In particular, critics of connectionism claim that it uses the *wrong kind of semantics for mental symbols*. Thus, in a famous paper, Fodor and Pylyshyn (1988) have suggested that the notion of *mental content* needed for cognitive symbol processing should be combinatorial and structure-sensitive. Any account that omits these features the way connectionism does is necessarily unable to explain some of the most salient aspects of cognition. This argument has released a surge of controversy over the semantic foundations of connectionist theory, which continues to the present day.²

The conceptual-semantic discussion on the status of connectionism represents the state of the art. I, for one, believe it is a very important discussion, though probably not for the same reasons as the main contributors to the debate. I think the discussion reveals aspects of *cognitive science generally*, irrespective of whether symbolism or connectionism is correct. In particular, it lays bare a presumption of methodological internalism, issuing in the idea that mental content is essentially an *intrinsic* property of a certain class of symbols. I believe this presumption to be misguided in important respects. In this chapter and the next, I follow its career in the context of connectionism. The notion of intrinsic content will be developed more fully in subsequent chapters.
Neural epistemics

I will use the following plan to achieve my aim in this chapter. Section 1 offers an informal introduction to some basic elements of connectionist architecture. In section 2, the main parameters of connectionist modeling are examined from a philosophical point of view. I then pause briefly to praise the cognitive power of connectionism, which holds the promise of a truly general neural epistemics, a “unified theory of the mind/brain”, as Patricia Churchland (1986) has called it (section 3). In the remaining sections, I discuss in some detail the semantic charges leveled against connectionism. Following up on the dual aspect theory of the propositional attitudes proposed in the previous chapter, I urge for the desirability and possibility of a conceptual interaction between folk psychology’s descriptive vocabulary and the explanatory resources made available by connectionism.

1. Elements of connectionist architecture

**Nodes and connections**

A canonical connectionist model is a network of interconnected processing units, loosely based on biological neurons. Following usage, I refer to these units as ‘nodes’. Nodes can be in various states of activation, usually represented as a real or discrete value between 0 and 1. In a very simple case, nodes are either on (1) or off (0). Signals (stimuli) are passed along connections between the nodes, comparable to the axons of nervous systems. These connections have characteristic efficiencies or ‘weights’ (the rough equivalent of synapses), usually depicted as values between -1 and 1. Negatively weighted connections are inhibitory: they tend to decrease the state of activation of the nodes onto which they project. Positively weighted connections are excitatory: they tend to increase the state of activation of the nodes onto which they project. If the activation state exceeds a critical threshold, the node will give off a signal to the next layer of nodes to which it is connected.

Depending on the nature of the connectivity between nodes, nets of different topology may be conceived. Some examples are shown in figure 4.1. To the left, two so-called ‘feedforward’ nets are pictured. In nets of this architecture, information flows in one direction only, going from input layer to output layer, while passing one, two, or more layers of connection weights. The nodes that are not directly visible to the outside world are traditionally called ‘hidden nodes’. They are involved in the internal processing of information. Despite their simple architecture, nets of this form are
capable of surprisingly complex cognitive performance, as we shall see shortly. The third example shows a so-called ‘recurrent’ net, which is quite different from the other two. Here the nodes have reciprocal connections, so that information not only flows forward, but may also be fed back to earlier stages of processing. Nodes may even latch onto themselves, thus tending to increase (or decrease, in the case of inhibitory connections) their own level of activity. This feature of autoconnectivity insures that, when a node is on, it will stay on, and when it is off, it will stay off.

*Elementary neural computations: feature detectors*

Although the mathematics involved in connectionist modeling tends to become highly complicated as various constraints are added, the basic principles are not difficult to understand (for some accessible introductions, see Caudill and Butler 1992; Jordan 1986). Three relevant levels of analysis may be distinguished here: that of individual nodes, that of nodes organized in layers, and that of the net as a whole. Characteristic computations are performed at each of these levels, displaying properties that are highly interesting from a cognitive point of view. I discuss each of these levels in order.
Consider first the level of the individual node. Figure 4.2 pictures a simple feedforward network consisting of three input nodes and a single output node. When the input nodes of this network are activated, they propagate a signal through the weighted connections. The level of activation of the output node is determined by the weighted sum of the incoming signals. This can be realized in a number of ways, but for present purposes I assume that the node computes the following function:

1. \( b = a_1 w_1 + a_2 w_2 + a_3 w_3 \)

where \( b \) and \( a_i \) designate levels of activation of the output and input nodes, and \( w_i \) designates the connection weight for the \( i \)th input node.

Generally, let \( a_i \) and \( b_j \) designate the levels of activation on nodes in contiguous layers, and let \( w_{ij} \) designate the weight of the connection from the \( i \)th node of the lower layer to the \( j \)th node of the upper layer. Then, for a total of \( n \) incoming connections, node \( j \) computes the following function:

2. \( b_j = \sum_{i=1}^{n} w_{ij} a_i \)
Although the operation described here is quite elementary, it is already very interesting from a cognitive point of view. To appreciate this, we need to consider the geometrical interpretation of the function computed. The input values and the weight values in the example may be represented as two vectors in three-dimensional Cartesian space, an input vector $\mathbf{a} = \langle a_1, a_2, a_3 \rangle$ and a weight vector $\mathbf{w} = \langle w_1, w_2, w_3 \rangle$. The value $b$ computed by the output node is known as the inner product (also called the scalar product or dot product) of these two vectors, defined as:

$$3. \langle a_1, a_2, a_3 \rangle \cdot \langle w_1, w_2, w_3 \rangle = a_1 w_1 + a_2 w_2 + a_3 w_3.$$ 

Generally, for two $n$-dimensional vectors the inner product is defined as

$$4. \mathbf{a} \cdot \mathbf{w} = a_1 w_1 + a_2 w_2 + \ldots + a_n w_n.$$ 

An alternative definition of the inner product of two vectors that is independent of coordinate systems and that admits an easy geometrical interpretation is

$$5. \mathbf{a} \cdot \mathbf{w} = |\mathbf{a}| |\mathbf{w}| \cos \phi,$$

where $\phi$ is the included angle between $\mathbf{a}$ and $\mathbf{w}$, and $|\mathbf{a}|$ and $|\mathbf{w}|$ are the lengths of the vectors.

As it turns out, the inner product of two vectors is a convenient way to measure how close one vector is to another. This is readily seen when we represent the vectors graphically in a vector space (figure 4.2, right). Elementary trigonometry shows that $|\mathbf{a}| \cos \phi$ is just the projection of $\mathbf{a}$ on $\mathbf{w}$. Hence, ceteris paribus, the larger the included angle, the smaller the inner product will be. Input vectors close to the weight vector evoke a positive response from the output node ($\cos 0 = 1$), those near an angle of $\frac{\pi}{2}$ tend to evoke a zero response ($\cos \frac{\pi}{2} = 0$), and vectors pointing in an opposite direction elicit a negative response ($\cos \pi = -1$). In point of fact, the individual node is operating as a feature detector, calculating how similar the input vector is to the resident weight vector. This property is obviously very interesting from a cognitive point of view. (Notice, however, that the device has very limited capabilities: in particular, it cannot discriminate between long input vectors at large angles and short vectors at small angles.)
Vector transformations and pattern mapping

Turning to the level of layers, we find a qualitatively new variety of computation, displaying features that are again most significant from a cognitive point of view. The simple model described above can be extended by adding more nodes to its input and output layers, as shown in figure 4.3. The extended net cannot only detect more features, as described above, it can also perform a new variety of tasks, generally referred to as pattern mapping, and including pattern recognition, pattern completion, pattern transformation, and pattern association. The importance of these tasks for perception, motor control and cognition has been stressed by many authors (see, for example, Dreyfus and Dreyfus 1986; Paul Churchland 1986 and 1990; Margolis 1987; Richards 1988; Churchland and Sejnowski 1992).

While individual nodes are computing scalar quantities, layers of nodes compute vectors: using the matrix of weights on the incoming connections, each layer transforms its input vector (representing an input pattern) into an output vector (representing the corresponding output pattern). For example, the output layer in figure 4.3 transforms a four-dimensional input vector \( \mathbf{a} \) into a three-dimensional output vector \( \mathbf{b} \):

\[
6. \ \langle b_1, b_2, b_3 \rangle = M \langle a_1, a_2, a_3, a_4 \rangle \\
= \langle w_{11}a_1 + w_{21}a_2 + w_{31}a_3 + w_{41}a_4, \\
w_{12}a_1 + w_{22}a_2 + w_{32}a_3 + w_{42}a_4, \\
w_{13}a_1 + w_{23}a_2 + w_{33}a_3 + w_{43}a_4 \rangle.
\]

\[\mathbf{b} = \langle b_1, b_2, b_3 \rangle\]

\[\mathbf{a} = \langle a_1, a_2, a_3, a_4 \rangle\]

Figure 4.3: Computational activity in layers of nodes

Layers of nodes perform vector transformations. In the example shown here, a four-dimensional input vector \( \mathbf{a} \) is ‘pushed through’ the connectivity matrix \( M \), and is transformed into a three-dimensional output vector \( \mathbf{b} \). This feature enables neural nets to perform a variety of cognitively interesting tasks known as pattern mapping.
In this equation, \( M \) is the matrix of weights on the incoming connections. Generally, the transformation of an \( i \)-dimensional vector into a \( j \)-dimensional output vector is given by the following equation:

\[
7. \langle b_1, b_2, \ldots, b_j \rangle = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1i} \\ w_{21} & w_{22} & \cdots & w_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} \end{pmatrix} \cdot \langle a_1, a_2, \ldots, a_i \rangle
\]

The mathematics of vector transformation is a powerful tool for describing the activity of nets. On the one hand, the matrix of connection weights determines both the structure of the net and the function it computes. On the other hand, it also enters into the description of the net’s layer-to-layer dynamics, that is, the way patterns of activation spread over the net as a function of time. Given a configuration of activity at \( t = i \), the net’s state at \( t = i + DT \) follows from its structural equations. I presently come back to the dynamics of neural networks.

**Linear and non-linear associators**

The architectures sketched so far have all been linear associators, mapping input to output in an indiscriminate, linear fashion. At the node level, I have been assuming that all nodes are computing linear functions from input signal to output signal. The limitations of this approach become particularly clear if we consider the function of multi-layer networks. Figure 4.4 shows a small multi-layer network model. An extra layer of connections and nodes has been added to our previous model, introducing a layer of hidden nodes \( b_i \). Now, the introduction of hidden nodes only makes sense if they are given non-linear rather than linear functions to compute. Otherwise, the same function can be computed with a single layer of connections, as discussed above. To enable nets to do something new by adding hidden nodes requires a non-linear architecture.

The function from a node’s input signal to its output signal is technically known as it’s transfer function. So far I have been assuming that the transfer function is just a one-step process: output signal = net weighted input. In most actual networks, however, it consists of three distinct steps. First, a node computes the weighted sum of incoming signals. Next, it computes its activation level as a function of this input. Finally, it calculates an output
signal from the activation value. While the first step is usually linear and follows equation (2) above, the second and third step typically bring in a non-linear component. In most networks, the function from input to activation level is a sigmoid curve. A commonly used sigmoid function for computing activation from input I is:

8. \( f(I) = \frac{1}{1+e^{-I}} \)

which has the useful property that its first derivative is very easy to calculate, namely:

9. \( \frac{df(I)}{dI} = f(I) \cdot (1 - f(I)) \)

The exact shape of the sigmoid is not particularly important, however; what matters is that it be monotonically increasing and bounded with both upper and lower limits. The final step from activation value to output signal typically uses a threshold bias. If \( f(I) \) fails to reach a predetermined value, the node computes zero as its output; otherwise, its output is \( f(I) \).

From a cognitive point of view, the addition of non-linear components to the transfer function has several advantages. In the first place, a threshold

\[
\begin{align*}
M_1 &= (a_1, a_2, a_3, a_4) \\
M_2 &= (b_1, b_2, b_3) \\
c &= (c_1, c_2, c_3, c_4)
\end{align*}
\]

**Figure 4.4: The contribution of hidden nodes**

Also multi-layer networks are engaged in vector transformations, pushing input vectors through consecutive banks of connection weights. The addition of hidden nodes in a network model only makes sense, however, if the nodes are computing non-linear transfer functions, such as the sigmoid curve pictured here (left).
bias in the net’s transfer functions enables the net to filter out any noise and other marginal activity that is irrelevant to the cognitive function to be computed. Linear associators, by contrast, suffer from the fact that they compute noise and nonsense along with useful information. Moreover, it is not easy to see how purely linear nets can generate information that is significantly more useful than what it was fed with in the first place. Suppose the input to a net is ‘barking’, ‘four legs’, and ‘tail’, then what we want to hear is ‘dog’. But if the output layer signals ‘dog’ iff the ‘barking’, ‘four legs’ and ‘tail’ nodes are all active, then why bother with the ‘dog’ node at all? We want a net to recognize a dog by its barking, legs, or tail alone, even if none of the other elements are perceived. Actually, this can be realized in a number of ways, even in simple nets. One important aspect of the relevant technique is coding a suitable prototype vector into the connection matrix (either directly, or as the outcome of a learning process). With the proper non-linear transfer functions on the nodes, the net will tend to produce an output that is as close as possible to this pre-stored vector given only part of it as input. This feature of vector completion is technically known as ‘autoassociative content-addressable memory’ (Hinton et al. 1986; Churchland and Sejnowski 1992, 80ff; Bechtel and Abrahamsen 1991, 62ff). Although it is not difficult to implement vector completion tasks for certain specific purposes, it is unclear whether there is a generally best technique, or which kinds of solutions may be best suited for domains and tasks as yet unexplored. An example of vector completion in a recurrent net will be discussed below.

Nonlinear nets can handle tasks that are much more complex than anything a simple linear net can do; they cannot only ‘look up’ the requested output in their connectivity matrix, but also correct the question and hypothesize upon its implications. This becomes especially clear when the dynamics of network processing are taken into consideration—the final computational feature to be mentioned here.

Beyond instant association

The nets discussed so far have all been more or less instant associators. In the early 1980s, it was discovered that nets can also be usefully studied from a dynamic perspective, as processes whose behavior is in many respects similar to that of thermodynamic systems. This made available the mathematical techniques applied in quantum physics for describing such systems, and led to the development of nets of new and more complex
architectures. I will pause briefly to discuss two such types of net here, Hopfield networks and Boltzmann machines. Figure 4.5 shows the trajectory in state space followed by a net seeking its lowest ‘energy state’. Shown in the inset is a 2D projection of the energy landscape traversed from starting-position (A) to desired position (E). The net is first ‘heated up’ by its input and then slowly ‘cooled down’. By applying the mathematics used for studying annealing phenomena in thermodynamics, physicist John Hopfield was able to devise a connectivity scheme that guides nets into stable states (Hopfield 1982). This type of net is known as a relaxation net.

The problem here is how to avoid that the system gets stuck in a local pocket such as B or D in figure 4.5. It can be shown mathematically that Hopfield nets are very good at finding good solutions rapidly, but that they are generally unable to find the best solution. In other words, they are unable to reach a global minimum such as E in figure 4.5. The problem can be solved, however, by a slight modification of the Hopfield rule. The original Hopfield net used a simple binary threshold rule for computing node activity: if the change in input activity (weighted sum of active input connections) is positive, the node computes 1 as its output; otherwise, the output is 0. Hinton and Sejnowski (1983) replaced this rule by a continuous sigmoid function, specifying the probability of output 1 as a function of change of

![Figure 4.5: Network dynamics](image)

Trajectory in state space of a Hopfield/Boltzmann net seeking its lowest energy state. The net’s ‘temperature’ is given as a function of the activation of two of its nodes. The inset shows a 2D projection of the energy landscape. B and D are local minima, as opposed to the global minimum in E.
input activity. By analogy to the thermodynamic systems studied by Ludwig Boltzmann in the nineteenth century, these nets are called Boltzmann machines (Hinton and Sejnowski 1983). Unlike Hopfield nets, they are guaranteed to settle into the global minimum, but only if they are ‘cooled down’ slowly enough. Which type of architecture best suits a given task is impossible to say in general; finding the best solution to a problem is obviously desirable, but sometimes speed will have priority over accuracy.

Figure 4.6 A simple relaxation net for perception of the Necker cube

The net shown here is massively recurrent; all internode connections are reciprocal. Inhibitory connections are marked by a bullet; all other connections are excitatory. Note that input and output connections have been suppressed in the diagram; in principle, each individual node can serve as input device as well as output device. Nodes are ordered in two groups of eight: \{A+, C-, B+, D-, E+, G-, F+, H-\} on the left, and \{A-, C+, B-, D+, E-, G+, F-, H+\} on the right, corresponding to the two geometrically consistent interpretations of the Necker cube. The net is intended only to illustrate certain abstract dynamic properties of connectionist models, without making any attempt at psychological or biological realism. (Adapted from Rumelhart and McClelland 1986a, vol. 2, 10.)
A simple relaxation net

The most difficult part of relaxation nets, from a cognitive point of view, is how to understand the cognitive significance of their ‘energy landscape’. In particular, what is the sense in which energy minima can be understood as ‘solutions’? To understand this point, it is important to see that a relaxation net is involved in constraint satisfaction. Given its initial state, in which it is ‘heated up’ by specific input values, the net seeks out the optimal distribution of activity as determined by its specific connection constraints. It is the weight matrix that defines the particular energy landscape in which the net finds itself. Now, if the matrix is carefully chosen so as to mirror the structure of a given task domain, the net will act as a problem solver in that domain. An example discussed by Rumelhart and McClelland (1986a, vol. 2, 8ff) is a net for the perception of ambiguous figures such as the Necker cube. Figure 4.6 shows part of this net. Each node stands for a particular hypothesis concerning the spatial location of the cube’s vertices. The node marked ‘H+’, for example, indicates that vertex H is located in the foreground, whereas ‘H−’ corresponds to the conjecture that it is located in the background. The net’s connectivity mirrors a number of semantic and epistemic relations that hold between the various hypotheses. If it is normally to be expected that the object of perception, when it has property $P_1$ (for example, edge CA pointing towards the observer), will also have property $P_2$ (for example, edge FH pointing away from the observer), then the connection between $P_1$ and $P_2$ nodes will be reciprocally excitatory; similarly for inhibitory relations.

The structural information built into the connectivity matrix is based on a prior geometrical analysis of the task domain. In addition, the net receives empirical information from outside. One or more nodes are activated by incoming signals (note that the input connections are not shown in the diagram). A positive input to a node reflects the probability that the corresponding hypothesis is correct; the stronger the input value, the more probable the hypothesis will be. Together, the ‘a priori’ connection weights and the ‘a posteriori’ input values determine how activation will spread over the network as a function of time. The net will try to reach a stable state, minimizing the number of simultaneously active, contradictory hypotheses, while maximizing the number of simultaneously active, mutually reinforcing hypotheses. Not surprisingly, the Necker net has two stable configurations of node activity, $\{A+, C-, B+, D-, E+, G-, F+, H-\}$ and $\{A-, C+, B-, D+, E-, G+, F-, H+\}$, corresponding to the two geometrically consistent interpretations of the Necker cube.
The Necker task exemplifies a whole range of problems that can be solved by the same net, provided their constraint structure is isomorphic to that described above. Moreover, the approach can be expanded to deal with less straightforward tasks. One possible addition is that of probabilistic or statistical constraints on the relation between hypotheses. For example, if the conditional probability of property \( P_1 \) relative to \( P_2 \) is small, this may be reflected in the connection weights \( w_{ij} \) holding between the pertinent nodes, which then may take smaller values \( |w_{ij}| < 1 \).

The above example also illustrates one possible way in which one or more preset vectors can be built into a connectivity matrix, as mentioned earlier in this section. The Necker net is engaged in a vector completion task: even if only part of the desired vector is given as input, say \( \langle A^+ \rangle \), it manages to reconstruct the full vector \( \langle A^+, C^-, B^+, D^-, E^+, G^-, F^+, H^- \rangle \). Moreover, the net is able to do much more than merely ‘look up’ the desired vector in its matrix. It considers competing hypotheses, chooses from among several stored vectors, and may even correct the input: for example, when it decides to suppress the contribution of \( \langle D^+ \rangle \) in input \( \langle A^+, D^+, H^- \rangle \).

2. A conceptual universe for connectionist theory

If the previous section described the ‘hard core’ of connectionism, in a loosely Lakatosian sense, I now turn to its ‘positive heuristics’, the latitude of choice available for constructing specific theory versions. Like ‘Turing machines’ and ‘Von Neumann machines’, the term ‘neural net’ carves out a vast conceptual space, in which specific models occupy only small regions. Analogous to the space of philosophy of mind described in chapter two, the philosophical latitude of connectionism can be represented in terms of a geometrical projection of its conceptual universe. I suggest three dimensions of freedom that are particularly relevant in this context: modularity, plasticity, and distribution of representations. As shown in figure 4.7, the result is a 3D universe, in which individual models are represented as vectors, and families of models as volumes.

The parameters adopted here are not unique for connectionist models. Also traditional cognitivist theories take specific degrees of modularity, plasticity and representational distribution, though arguably not always in the same value range as connectionist models. Nor are the three parameters the only dimensions of connectionist freedom. Other aspects may include the degree of biological realism (the extent to which the model matches structure and performance of nervous systems), and the degree of
eliminativism (the extent to which the model defies folk psychology). These other parameters generate a conceptual universe of higher dimension, of which figure 4.7 shows only a 3D projection. Finally, notice that the dimensions are assumed to be conceptually independent of one another. This is probably a simplification of the truth. In the following paragraphs, I shall make one or two remarks to the effect that the dimensions may be interconnected in important ways.

**Plasticity and learning**

One of the most intensively studied features of connectionist nets is their capacity for autopoiesis or self-organization, the ability to change their internal structure and to learn to instantiate new cognitive functions. The net’s plasticity is inversely proportional to its structural rigidity; the more connection weights and threshold biases tend to be fixed, the less capable of learning new i/o functions the net will be. Two extreme positions can be distinguished here, known traditionally as nativism and empiricism. According to the nativist version of connectionist modeling (located on plane AEHD in figure 4.7), the net’s structure is determined at birth—literally so, if the net serves as a model of the human mind, and figuratively speaking,

![Figure 4.7: Part of the conceptual universe of connectionist theory](image)

Three dimensions of freedom for connectionist modeling are shown, resulting in a 3D projection of its conceptual universe. For convenience of exposition, it is assumed that the dimensions are conceptually independent of one another, and can be represented as orthogonal axes.
if it is a purely pragmatic model in commercial AI. The connectivity structure is fixed by nature or by man, so as to implement the desired cognitive function. By contrast, according to the empiricist version (located on plane BFGC), the net is initially a tabula rasa, a ‘universal’ system that knows nothing but that can learn anything. Only as a result of learning processes, it develops the structure for instantiating the required cognitive function. In seventeenth-century philosophy, nativism and empiricism were virtually irreconcilable positions, as is testified by Locke’s criticism of Descartes (Locke 1670, book I). Today, connectionism opens up the fascinating, and empirically testable, possibility of bridging the philosophical gap with a fine-grained mix of malleability and rigidity (cf. Bechtel and Abrahamsen 1991, 101ff).

As we saw in the previous section, the connectivity structure of neural

---

**Figure 4.8: Part of a neural net for learning the past tense of English verbs**

The diagram shows a feedforward net with two layers of hidden nodes, involved in transforming the root ‘run’ to its irregular past tense ‘ran’. For pictorial ease, the net has been simplified in several respects; important corrections will be discussed below. (Adapted from Rumelhart and McClelland 1986a, vol. 2, 222.)
nets determines the cognitive function that is computed. For the performance of a net it makes no difference whether this structure is acquired or innately specified, unless, of course, the structure is very difficult to acquire, or takes an unacceptably long time to acquire. Rigidity can be a valuable asset. Consider the Necker net described above, in which the weight matrix was preset to match the a priori structure of the task domain. Although for the researcher this meant some extra work, the net can immediately profit from this. Translated in terms of biological nets, organisms may profit from the cognitive heritage of previous generations, when structure is accumulated in a phylogenetic rather than an ontogenetic learning process. Realistic examples of this are given by Marr and others in their work on early vision (Marr 1982).6

In addition to analysis of the task domain, study of the nervous system can be of service here. An understanding of the actual connectivity of specific parts of the nervous system, engaged in solving specific problems, may reveal important aspects about the structure of the cognitive domain that are otherwise hard to come by. A mere top-down analysis of cognitive tasks may be insufficient for determining their relevant structure: many different algorithmic procedures may be equally able to solve a given problem in the absence of specific constraints on the procedures’ feasibility (memory resources, processing speed, hardware implementation, interaction with other routines, etc.). In this regard, the search for innate connectivity structure is a clear example of conceptual interaction between top-down and bottom-up approaches, as urged in the previous chapter.

In practice, research in connectionist modeling tends to concentrate on the learning capacities of empiricist nets. A typical example is the model for past tense formation of English verbs, proposed by Rumelhart and McClelland (1986a, vol. 2, 216-271). Part of the net in question is shown in figure 4.8. An array of input nodes, connected to an ‘ear’, is fed with phonetic representations of verb forms. An association network, consisting of two layers of hidden units, transforms this input first to its root form, which is then associated with the corresponding past tense. Finally, an array of output nodes connected to a ‘mouth’ transform the result into a phonetic output. The interesting point of this net is that part of the appropriate structure (the connections between the layers of hidden nodes) was not built in but acquired. Starting from a random configuration of connection weights, the net was subjected to several hundreds of trial runs. At each cycle, a verb was presented to the net, and the difference between the actual output and the desired output was used to slightly adjust the connection weights ac-
According to a standing algorithm. After training, the net performed admirably well, and was able to go beyond the stock of verbs on which it had been trained with a success rate of about 90%. Apparently, the net had managed to latch onto the rule for forming the past tense of regular verbs, as well as memorized the relevant exceptions to the rule. Interestingly, it showed a characteristic learning profile. At first, irregular forms such as ‘run’/‘ran’ were learned correctly, but they were then overregularized into forms such as ‘run’/‘runned’. Eventually, the net settled on the correct forms again. This phenomenon is also characteristic of one phase of language learning in children.

Although the net described here may not be correct as a model for acquiring past tense formation (for severe criticism, see Pinker and Prince 1988; some replies are summarized in Bechtel and Abrahamsen 1991, chs. 6 and 7), it illustrates the key claims of connectionist learning theory: that learning occurs by incremental adjustment of graded parameters (connection weights), and that it is driven by the computed difference between implicit predictions and observed events (figure 4.9). A number of weight-adjusting procedures have been devised (for an overview, see Hinton 1989; 1986).

Figure 4.9: Learning as gradual descent in weight/error space
Connectionist learning occurs by incremental adjustment of graded parameters (connection weights), driven by the computed difference between implicit predictions and observed events. If the net’s performance (success rate) is given as a function of its connection weights, a characteristic learning trajectory emerges. In a sense, the above diagram shows the behavioral or external counterpart of the net’s internal dynamics pictured in figure 4.5. (Adapted from Churchland 1989a, 201.)
Churchland and Sejnowski 1992, ch. 3). An example is the so-called backpropagation algorithm, a descendant of the ‘delta’-rule used by Rumelhart and McClelland in the example described above. In the ‘backprop’ algorithm, as it is usually called, the difference between the net’s actual output and its required output, as defined by some external source or teacher, affords a measure of error. If there is a discrepancy, the strength of each connection to an output unit is adjusted in the direction, excitatory or inhibitory, that reduces the error, and in proportion to its effectiveness to do so. A similar process of adjustment occurs at the next layer down. In general, the contribution that a hidden unit makes to the overall error depends on its level of activation, and its strengths of connection to each unit at the next level up, as well as on those units’ contribution to the error (Hinton 1992; Johnson-Laird 1988, 187ff).

Due to their built-in power of plasticity, connectionist models are, generally speaking, better able to explain phenomena of learning than traditional accounts. Yet, there is a price to this capacity. I mention three drawbacks here: training, statistics, and opacity. First, huge numbers of training runs are needed in most cases before the net is fully operational. Moreover, the net can discover the hidden invariants or rules only if the sample set contains a relatively large amount of regular cases. Though statistical learning is doubtless important for inductive generalization, it is not the only relevant factor. Human beings are often able to generalize upon a phenomenon after as little as two or three examples—a fact that cannot be explained by mere statistical association. Finally, the structure acquired after training tends to be cognitively opaque. Although the net is demonstrably capable of instantiating a given cognitive function, it is often difficult to see how it does so. Especially in large nets, consisting of many layers of hidden nodes and multiple connections between them, it is almost impossible to analyze the acquired weight matrix in terms of the structure of the cognitive problem solved by it. It has been remarked that radically ‘empiricist’ models are as useless as a map at real scale. ‘Nativist’ models, by contrast, suffer from the opposite snag: they depend on a prior analysis of the cognitive domain by the researcher. There seems to be a trade-off between learning and enlightenment here. If structure is built into the net, we do not understand learning, but if all knowledge is learned, it is difficult to analyze its structure. Obviously, both are equally necessary: structure is needed to guide learning, and learning is needed to advance structure.7
Chapter four

Modularity and functional holism

The idea of task-oriented modeling is no less appealing in connectionism than it is elsewhere in cognitive science. The view that a cognitive system makes use of one or more sets of relatively independent subsystems, each responsible for a limited subdomain of cognition, creates a latitude of choice for connectionist theory. Two gross options present themselves, one thoroughly holistic, the other thoroughly modular. In figure 4.7, these positions are located on planes DHCG and AEFB, respectively. On the panholistic view, our cognitive machinery is a giant ensemble of nodes and connections, with no cognitively relevant distinctions between different parts of the net. This view, for which it is difficult to find support, is in fact a connectionist version of Karl Lashley’s principle of ‘mass action’ or ‘equipotentiality’, discussed in chapter two (cf. Gardner 1985, 260ff; Van Gelder 1991, 48ff). According to modular versions of connectionism, on the other hand, specific functions are subserved by specific nets. This view is now commonplace. As for the question which theory will eventually turn out to be right, holism or modularity, the truth will probably be somewhere in the middle, a mere footnote to the history of science.8

In practice, all connectionist modeling is thoroughly modular. The cognitive functions nets are trained or built to instantiate are, without exception, highly specific and carefully defined tasks (cf. Murre et al. 1989). In chapter two we saw that modularity is liable to be a by-product of this method of research. If it is asked, for example, how the net for past tense formation fits into the general architecture of the language processor, the answer will typically expand on the way in which the net, considered as a separate entity, communicates with other nets subserving similar subtasks. Even if the requirement proposed by Fodor and others, that modules are ‘informationally encapsulated’, would be softened, so as to enable subnets to receive feedback from other nets, these subnets would still be acting as separate entities, specialized in the computation of domain-specific cognitive functions.

There are important connections between modularity and the question of plasticity. As was pointed out above, the acquired or innate connectivity structure determines which cognitive function a net is able to compute. This in turn automatically brands the net as an autonomously acting, cognitive module. Moreover, the power of plasticity exhibited by neural nets opens up the possibility of a battery of initially similar, modular nets that, in the course of a continuous process of learning and specialization, are put to work on different cognitive tasks, depending on their place in the overall
functional organization. Similar connections hold between modularity, plasticity, and questions of biological realism. Neuropsychological syndrome analysis may shed some light on the gross functional architecture of the brain, as will our understanding of specific neural pathways, such as the optic tract. Moreover, study of the brain’s more fine-grained modular organization, such as found in the laminar structure of the cortex, may provide connectionists with the information necessary for devising relevantly structured networks. Finally, theories of ontogenetic and phylogenetic neural plasticity, including neurophysiological evidence on learning processes, may be consulted for corroboration of connectionist accounts of learning in terms of incremental adjustment of graded connection weights. The conceptual proximity of connectionism and neuroscience gives us reason to expect that research in these fields may benefit from intensive cooperation.

*Distribution of representation*

If we want to know what the net is doing, and how it is doing so, we must necessarily turn to the semantic interpretation of node activity and connectivity. Generally speaking, we need an interpretation of vectors representing node activity, and of weight matrices representing connectivity. In particular, we expect the ‘built-in’ vectors of the weight matrix to contain a representation of the cognitive domain. Two extreme positions can be distinguished here, a localist strategy (located on plane ABCD in figure 4.7) and a distribution strategy (located on plane EFGH). In localist models, specific macroscopic features of the cognitive domain are represented by individual nodes. In distributed models, several nodes at once partake in the representation of a single feature. From a vectorial point of view, the difference between these strategies lies in the dimensionality of the representing vector: in localist models, individual categories are coded by unidimensional vectors, whereas distributed models use vectors of higher dimensions to this end.

In practice, multidimensional vectors are of the essence for connectionist modeling—hence the ‘D’ in ‘PDP’. They introduce the notion of ‘microfeature’ coding: if vectors represent macroscopic properties of the cognitive domain, then their components must represent microfeatures (Hinton et al. 1986; Smolensky 1987, 1988). The idea is best illustrated by means of a concrete example. In the past tense net shown in figure 4.8, I conveniently assumed that individual hidden nodes represent letters. In reality, however, the situation is much more complex. Each of the hidden nodes represents a so-called Wickelfeature, an adaptation of the phoneme coding system de-
vised by Wickelgren (1969). Thus, both hidden layers consist of a pool of several hundreds of nodes, each of which represents a different Wickelfeature. The past tense ‘came’ (/keim/), for example, corresponds to the ‘Wickelphone’ triplet \((\_K_{a'} \_K_{m'} \_M_{a})\), where ‘#’ indicates a word boundary, and each element takes its predecessor and successor as subscripts. Generally, each phonemic segment has been sensitized to its immediate context. In the actual net, these Wickelphones are translated into Wickelfeatures (and back again) by the fixed parts of the architecture. To be more specific, let us take \((K_{m'})\) as an example, the Wickelphone that targets on the vowel in ‘came’. We find that it is coded by sixteen Wickelfeatures, each a triplet such as (stop, low, nasal) or (unvoiced, low, voiced). Similarly, to represent the root ‘come’ (/kum/), 48 out of 460 different Wickelfeatures will be activated in the first hidden layer, 16 for each of its three phonemes (Rumelhart and McClelland 1986a, vol. 1, 233ff; Bechtel and Abrahamsen 1991, 178ff). The point of this example is that different aspects of the cognitive task are represented at a much more fine-grained level of analysis than figure 4.8 led us to expect; the grammatical regularities discovered by the net are captured by interactions among microfeatures rather than among macroscopic objects such as verbs and letters.

The representational issue is not always as straightforward. Consider the Necker cube network discussed earlier. Should vectors \(\langle A+, C-, B+, D-, E+, G-, F+, H-\rangle\) and \(\langle A-, C+, B-, D+, E-, G+, F-, H+\rangle\) be seen as distributed representations of two cube orientations, or rather as sets of local representations of the cube’s vertices? Similarly, does \(\langle A+, C-, B+, D-, E+, G-, F+, H-, A-, C+, B-, D+, E-, G+, F-, H+\rangle\) represent a ‘supercube’ of indifferent orientation? Is \(\langle A+\)\) the local and macroscopic representation of a vertex pointing toward the observer, or does it only make sense as a microfeatural part of a larger distributed representation? Is \(\langle A+, A-\rangle\) a representation of a ‘supervertex’ of indifferent position? Questions such as these defy explanation until the exact meaning of the terms ‘local’ and ‘distributed’ in this context has been clarified. Rather than attempting to give a general answer here, I propose the following rough and practical solution, which for the moment should suffice (I will come back to this issue in more detail):

10. If a vector \(\langle v_1, v_2, \ldots, v_n\rangle\), or a vectorial component \(\langle v_i\rangle\), can make no difference to the net’s overt performance, it is cognitively irrelevant. Hence, it cannot be meaningfully interpreted in terms of either macro- or microfeatures of the cognitive domain.
11. Features of the cognitive domain are macroscopic, relative to a net N, if they can be individually detected by N as demonstrated by its overt output.

12. Macroscopic properties may be composed of microfeatures.

13. Properties are microfeatural, relative to a net N, if they cannot be individually detected in N’s overt output, but turn up only at the level of input nodes and/or hidden nodes.

As intimated earlier, it is not always easy to determine which cognitive structure a net should use, or which structure it has in fact acquired as a result of learning. In the present context, this problem takes the form of the opaque structure of microfeatures. There is a connection with the nativism/empiricism parameter: if microfeatures are built into the weight matrix, their contribution is, by that token, also ‘semantically transparent’ (borrowing a term from Clark 1989). But if a cluster of nodes comes to represent a macroscopic property as a result of learning, it is not always easy to extract the microfeatural contribution made by individual nodes and connections. Various techniques for interpreting hidden node activity have been developed. One method, known as hierarchical cluster analysis, tracks the average activity levels on the hidden nodes as specific inputs are presented, thus establishing a correlation between input and node clusters (Sejnowski and Rosenberg 1987). Another technique makes use of so-called ‘Hinton diagrams’, in which the weights on connections from input units to individual hidden units are graphically displayed (figure 4.10). By comparing the weight configurations on different hidden nodes, some understanding of their representational function may be gained.

As a final link between the parameters discussed in this section, I introduce the notion of ‘micro-modularity’ at the subnet level, even as basic as the node level. All nodes and clusters of nodes are involved in computing specific functions, whether the net adopts a localist strategy or is massively distributive. To interpret the net’s activity as cognitive, is to see it as semantically evaluable, while in order to understand its semantics, we must turn to its internal structure. This in turn requires that we look for an interpretation of the vector transformations performed by the hidden nodes, either taken individually or considered as clusters. Hence, the net’s components inevitably function as micromodules whose autonomous computations subserve specific cognitive (sub)functions. In sum, if (clusters of) hidden nodes are not micromodules, then nets are not modules. This result is hardly surprising: ‘modularity’ is a functional notion, and as such invites
the idea of hierarchical organization and relations of dependency across levels.

3. Neural epistemics: the cognitive allure of connectionism

*Explaining biological intelligence*

Before turning to criticism, I want to pause briefly to praise the power of connectionism. The unique possibility opened up by connectionist models is that they may lead us to a realistic understanding of how biological intelligence works. For various reasons, neurobiology alone cannot accomplish this feat. One of these reasons we saw in the previous chapter: the cognitive puzzles solved by the brain are too complex to be analyzed by the conceptual apparatus of biology. Conversely, the biological complexity of the brain defies analysis in cognitive terms. Connectionism may bridge the gap between the two complexities. Its mathematical description of the dynamics of complex neural nets provides a link between, on the one hand, the topology of real nervous systems, and, on the other hand, the cognitive functions computed by specific network structures.

Connectionist models account for many of the properties that are typically displayed by biological cognitive systems. In addition to the obvious topological similarities between connectionist nets and nervous systems,

![Diagram](image)

*Figure 4.10: Hinton diagrams of weights on hidden nodes*

The figure shows Hinton diagrams for two hidden nodes in a fictitious net. Each node receives stimuli from five input nodes. The weights on the connections are rendered as squares, proportional in size to the value of the weights; white squares indicate excitatory connections, black squares mark inhibitory connections. The input nodes are fed with specific properties, such as ‘tail’, ‘barks’, or ‘furry’. On the basis of the weight matrix between input layer and hidden nodes (if not on output performance, which is here assumed to be absent), it seems reasonable to surmise that the node on the left represents the concept DOG, whereas the one on the right is a CAT categorizer. Apparently, the net believes dogs to be less furry than cats, while cats are more tailly than dogs. (See also Hinton and Sejnowski 1986; Lloyd 1989, 110-112.)
and the resemblances between the conceptual frameworks used for describing them, I mention seven strong points of connectionism.

14. *Graceful degradation.* Neural nets are relatively insensitive to damage to their parts. Because ‘data’ as well as ‘program’ are distributed over the net as patterns of connectivity, it will continue to perform adequately even if parts of it are impaired. The more nodes and connections are damaged, the less adequate its performance will be. This gradation of effect gives connectionist nets both neural and evolutionary plausibility.

15. *Robustness.* Connectionist models, like brains, are relatively insensitive to missing or erroneous data. They will function reasonably well even if the input or stored data are suboptimal for finding the solution to the cognitive problem. Again, the effect is graded.

16. *Swiftness.* In a typical net, the output may be virtually instantaneous as well as continuous. As soon as the input nodes are activated, the entire net is ‘heated up’; an output configuration is available almost immediately. This also applies to relaxation nets, in which a (probably suboptimal) output can be extracted at every stage. Neural nets obey the ‘one hundred step rule’ (Feldman and Ballard 1982): typical cognitive tasks such as word recognition are performed by the brain in tenths of a second; given the fact that each neuron takes several milliseconds to fire, these tasks can therefore require no more than roughly one hundred steps—much less than typically required by traditional serial accounts.

17. *Plasticity and learning capacity.* Due to their built-in power of plasticity, connectionist models readily account for cognitive development.¹¹

18. *Satisfaction of soft constraints.* The dynamics of relaxation enables nets to find the best overall solution to multiple constraints, even if this solution is incompatible with each of the individual constraints. This feature enables nets, inter alia, to extend old solution procedures to new problems.

19. *Content-addressable memory.* Network models are able to retrieve the same information from a variety of different cues that are part of the contents of the memory itself. Hence, there is no need for explicitly coded storing and look-up procedures. In addition to achieving computational parsimony, an important asset of this feat is its similarity to the phenomenology of human associative memory.

20. *Vector coding.* All of the above points converge on the notion of vector
coding, which offers a general account of the contribution made by individual nodes and connections (neurons and synapses) to the overall product. At the same time, it makes available the powerful mathematical device of vector transformations, and new physical metaphors for understanding thought (vector dynamics, graded relaxation).

Each of the above properties represents features we think of as being natural, in the loose but important sense of properties that we, living human organisms, possess as a matter of fact. Yet, traditional accounts find it very hard to explain these very properties. They can be built into classical models, but then they are designed to look natural—which is not how we intuitively conceive of them. Modifying a metaphor of Gould and Lewontin (1979) to fit my purpose, one might say that psychologists have been painting frescoes on the spandrels of Saint Mark’s, rather than concentrating on the spandrels themselves. The spandrels have a natural, architectural necessity as parts of the cathedral’s vaulted dome; the frescoes, by contrast, are beautiful but irrelevant.

**Neural epistemics: such stuff as dreams are made on**

Far more important than the fact that connectionism can explain specific characteristics of biological intelligence, is the fact that it does so in a coherent way. For the first time in the history of science, we have an approach that brings it within our reach to understand the buzzing complexity of the brain from a coherently cognitive point of view. The computational mechanism of vectorial dynamics is powerful enough to be applied to cognitive phenomena generally; it opens up the possibility of a unified theory of neural epistemics. Although Shakespeare would not have dreamt that his words might be used in this context (which, incidentally, is true of most contexts), neural nets are literally “such stuff as dreams are made on”. To drive this point home, I review some of the epistemically most salient aspects of connectionist computation.

First, the key notions of vector coding and vector transformation call attention to the way brains manage to meaningfully interact with reality. In organic nets, vectors are pushed through banks of living matrices of synaptic weights. Although the computations performed by individual neurons may be senseless in themselves, the vectorial transformations in which they partake are laden with cognitive meaning. At the input side, this explains in outline how the brain is able to transform the afferent sensory information into representations that are cognitively more useful to the or-
ganism. It manages to impose order and stability on the flux of ambient energy to which the organism is constantly exposed, extracting relevant regularities, learning and storing new categories, combining them with old information, and bringing them to bear on new. Similarly, at the output side, the same process explains how the organism’s internal representations are transformed into motor control of the appropriate proximal and distal efficacy.

Vector coding reveals the enormous representational capacity and processing power of the brain. The typical, adult human brain contains between 100 and 1000 billion neurons. Assuming that each neuron functions as a binary node, the brain can be in either of \(2^{100,000,000,000}\) different activation states (in the order of 10 to the power \(10^{10}\)). Let me call these unimaginably complex, global states supervectors, to highlight the fact that each of them can be a cognitively significant representation. Their number defies human imagination. Moreover, the global activation state changes each fraction of a second. Spontaneous and induced neuron activity, involving spiking frequencies from 1 to 500 Hz, cause the brain’s supervector to be vibrating at, let us say, 100 Hz.

If the above calculation gives the number of different possible supervectors, how many different supermatrices (configurations of synaptic weights) can be set to work on them? Each neuron connects with several hundreds or thousands of other neurons (4-5,000 on an average motor neuron, against 90,000 on a single Purkinje cell in the cerebellar cortex). Hence, the total number of neural connections in the human brain is of the order of \(10^{12} \times 10^3 = 10^{15}\). Assuming (not unrealistically) that each connection weight can take ten discrete values, this buys us a perplexing number of possible brain configurations, ten for each of the \(10^{15}\) connections, or ten to the power \(10^{15} = 10^{1,000,000,000,000,000}\). Suppose that only one percent of these configurations can be realized, for whatever reason, and that 99.99 percent of all really possible configurations make no sense at all. This still leaves us with an incomprehensibly large amount of cognitive plasticity: \(10^{1,000,000,000}\) different, cognitively relevant supermatrices. Recall that the total number of elementary particles in the entire universe has been estimated to be a mere \(10^{87}\) (Churchland 1989a, 131-132, 189-190, 209-210).

What are all these vectors and matrices doing? An important facet of connectionist nets is vector completion, which, as we have seen, serves as a mechanism for ampliative inference. It explains how nets are able to enhance or enrich the information available in the input vector alone. The added information derives from the weight matrix, which may contain one
or more prestored vectors to which the input vectors are computationally assimilated. The relevant connectivity structure may be either learned or innately specified. The task of vector completion is less trivial than it may seem: the weight matrix need not contain the answer to specific questions (although it may do so); rather, it contains the means for solving general types of problems. This sets it apart from a simple look-up table, as explained earlier.

The idea that vector coding is relevant for non-trivial processing tasks can be graphically expressed in terms of phase space partitioning and prototype activation. Figure 4.11 (top left) shows the distribution of activity on ten hidden nodes that turns out to have the same effect as a given ‘prototype’ vector (bottom left). As the original vector is fed into the weights con-

![Diagram](image)

**Figure 4.11** (left): Activation vector assimilated to prototype vector

The diagram shows a vector consisting of ten components (dimensions), each representing the activity level of a hidden node. (Note that activation states are assumed to be continuous rather than binary.) The original activation vector (top), though differing from the prototype pattern in several respects, as marked by the curve, is assimilated to a prototype resident in the connectivity matrix (bottom). The net acts as essentially a curve-fitter, interpolating new values between those that remain unchanged (cf. Churchland and Sejnowski 1992, 105ff).

**Figure 4.12** (right): Concepts as partitions on phase space

State space representation of activation levels on hidden nodes. Activation levels of only three hidden units are shown (against ten in the previous diagram). The state space is partitioned into two subvolumes separated by a ‘curtain’ or hypersurface. The partitions are defined as sets of vectors whose computational effects, *modulo* the weight matrix, are in relevant respects indistinguishable from those of a given (range of) prototype vector(s), shaded in the diagram. (Adapted from Churchland 1989a, 203. See also the ‘goodness-of-fit’ function for the recursive Necker net discussed in Rumelhart and McClelland 1986a, vol. 2, 8ff).
necting it to the next layer up, it is tentatively corrected by the weight matrix, so as to match the matrix’s resident prototype vector. This process of computational assimilation allows us to define a phase space partition as the set of vectors that compute relevantly the same outcome as a given prototype. Figure 4.12 shows two such partitions on a 3D vector space. The corresponding prototype vectors are rendered as volumes. If a vector falls on one side of the curtain, it is relevantly assimilated to a vector in the prototype region. Mathematically, the approach can deal with arbitrarily many partitions on spaces of arbitrary dimensionality.

I want to emphasize the point that partitioning is only a mathematical description of differences in computational efficacy. (This will be relevant for my discussion of connectionist semantics in the next chapter.) It is the weight matrix that defines the prototype. If a vector is assimilated to a vector in the prototype region of figure 4.12, its computational effects, modulo the weight matrix, are in relevant respects indistinguishable from those of the prototype vector.

From a cognitive point of view, the vector approach to representational dynamics is extremely fertile. In several respects, the model of prototype activation may offer a viable alternative to traditional, language-based models of cognition (Churchland 1989b; 1992a; 1992b). As an example, I mention the important phenomena of perception, categorization, and explanation. On the traditional view, these functions are subserved by entirely different cognitive mechanisms. In rough outline, processing of sensory input first forms stable percepts or ‘observations’. Next, these observations are categorized so as to yield observation sentences. Finally, explanation is achieved by subsumption of the observation sentences under a pertinent covering law. On the connectionist account, by contrast, these apparently dissimilar functions are subserved by a single mechanism, that of prototype activation. The ‘neural subsumption’ of input vector under prototype gives us all at once: construction of stable percept, activation of relevant concept, as well as explanation and prediction of unperceived aspects of the object by inference from the prototype. The type of coherent computational understanding exemplified here is what I refer to as ‘neural epistemics’.

Neural epistemics does not end with percepts and concepts. The approach can be extended to deal with other entities familiar to us from folk psychology. Supra-conceptual constellations such as theories and paradigms, processes such as explanation and inference, as well as ‘high-level’ features of scientific development, are all equally within the reach of this speculative expansion. For example, why is simplicity a virtue of theories...
and of theoretical explanations? Seen from the neuroepistemic perspective, a very natural answer suggests itself. Theories that involve too many hidden units tend to frustrate the learning process by settling on ad hoc solutions for each sample of the training set, rather than generalizing upon the set as a whole. Unexpectedly, we see that simplicity does not fight the vice of complexity, but that of being ad hoc (Churchland 1989b, 83ff).

In a similar vein, Churchland has recently argued for the vindication of some well-known themes from the philosophy of Paul Feyerabend, including theory-ladenness of observation, incommensurability of competing theories, and proliferation of theories as well as of methodologies (Churchland 1992a; cf. Feyerabend 1962 and 1975). Theory-ladenness, for example, is readily explained in terms of a distinction between mere peripheral transduction (“Eye-balls and cameras are blind”, as Hanson (1958) inculcated on us), and observation as a cognitive achievement performed by banks of synapses laden with prestored vectors. Moreover, the incommensurability of large-scale, competing theories starts to make sense if we consider the fact that these theories are alternative configurations of the same populations of synaptic weights; hence, learning a new theory necessarily implies that old evidential biases (connection weights) are destroyed.

As a final example of the kind of high-level explanations neural epistemics may have in store, I call attention to instant learning. How to account for sudden flashes of insight, serendipity, and conceptual breakthroughs that play such an important role in everyday life as well as in science? At first sight, the phenomenon is more readily explained in terms of traditional models. Mental items such as rules and concepts can be replaced instantly; the application of a new concept, or of an old concept to a new domain, can account for the sudden increase of a theory’s explanatory and predictive success. Learning in connectionist nets, by contrast, typically proceeds by slow and continuous adjustment of graded weights; there is no apparent apparatus for instant conceptual change. Yet, connectionism can deal with this problem. Churchland has suggested a plausible model for conceptual redeployment, which enables nets to instantly improve their performance (Churchland 1989a, 236-243). As an example, Churchland mentions Huygens’s breakthrough in optics, which was achieved by bringing to bear on optics certain well-known facts of wave theory, which had previously been confined to the study of mechanical waves such as sound and water waves.

“There was no need for Huygens to effect a global reconfiguration of his synaptic weights to achieve this conceptual shift. He had only to appre-
hend a familiar class of phenomena in a new cognitive context, one supplied largely by himself, in order to have the inputs activate vectors in an area of his conceptual space quite different from the areas they had previously activated. The difference was the context fixers brought to the problem. (…) The novelty (…) consisted in the unusual redeployment of old resources, not the creation of new ones. No new resources were created; nor were any old resources destroyed” (Churchland 1989a, 237-238).

The process of conceptual redeployment in neural nets is schematically shown in figure 4.13. Standard connectionist learning of the kind described earlier is not only slow, but inevitably destroys all previous concepts (partitions on activation space) stored in the old weight matrix. A discontinuous process like the one described here will not only be very fast, but also will leave the old weight matrix largely intact, thus conserving previously acquired concepts. Of course, the process is still a redistribution of weighted connections, but not of connections within the net itself. Rather, it must be internet connections that are formed or readjusted, such that the input to net

![Diagram of neural network](image-url)

*Figure 4.13: Conceptual redeployment*

The figure shows the ‘vector fit’ (success ratio) of a given, modular net N as a function of connection weights on its hidden nodes, only two of which are taken into account here. No graded improvement of performance takes place, but rather a sudden, disruptive rise. After fruitless meanderings from A to B, the net gains access to other resources, external to the modular net itself. The dramatic increase of vector fit is still the result of a redistribution of weighted connections, but obviously not of $w_i$ or $w_j$ (nor of any of the other internal connections in N).
N is led to a different net M, where it is processed by M’s resident phase space partition, and then fed back to N again. These connections serve as the ‘context fixers’ referred to in the above quotation; they enable N to profit from the computational resources of a different net M.

Notice that Churchland’s redeployment model implicitly assumes that theories are relatively autonomous, domain-specific nets, which act as cognitive modules. An interesting consequence of this is that learning phenomena, and instant learning in particular, may be relevant for our understanding of the modular organization of neural nets. If graded (parameter-) learning is largely an effect of the ‘internal’ plasticity of modular nets, instant learning involves plasticity of a different order, namely, of intermodular rather than intramodular connections. It is an empirical question whether these suggestions are correct. Once more, they highlight the fact that the various dimensions of connectionism’s conceptual universe, as described earlier in this chapter, are interrelated in important ways.

Whither folk psychology?

Is connectionism a threat to folk psychology? I do not think so. The conservative picture sketched in the previous chapter, hinging on the idea of conceptual interaction between folk psychology and cognitive science, is borne out by our examination of connectionism. At almost every stage of the foregoing discussion, we found top-down constraints derived from folk psychology weighing heavily on connectionism. I briefly rehearse some of the most pertinent results:

21. Modeling vector completion tasks requires selection of prototype vectors, based on a prior understanding of the cognitive domain.

22. In connection with this, prior analysis of the task domain is needed for determining the innate connectivity of relaxation nets.

23. Conversely, the acquired structure in empiricist nets tends to be cognitively and semantically opaque. Methods for extracting structure from the connectivity matrix, such as the method of Hinton diagrams, are essentially hermeneutical, and rely largely on considerations of top-down plausibility.

24. Generally speaking, network activity can be viewed from two different perspectives: as interpreted and as uninterpreted vector dynamics. In their uninterpreted form, vector transformations are merely abstract, mathematical descriptions of neural activity, without consideration of the cognitive function they may be subserving. In their interpreted
form, they are revealed as mirroring the relevant structure of a given task domain.

25. The cognitive allure of neural epistemics lies in its ability to vindicate and explain sophisticated phenomena such as theory change, explanation, conceptual breakthrough, as well as learning, perception, and categorization. This means that these phenomena have already been specified as explananda in folk psychology, and that these explananda have been *adopted* by connectionism.

26. Finally, we should not forget that even elementary features such as automatic noise reduction, robustness, swiftness, and content-addressable memory are cognitively desirable by virtue of the fact that they vindicate aspects of phenomenological psychology and anthropology. This tends to be obscured by the fact that the features in question follow quite naturally from intrinsic properties of simple nets. Their *naturalness*, however, does not in any way diminish their *cognitive significance*.

The first three points, (21)-(23), draw attention to the fact that connectionism, when considered as an isolated bottom-up approach, suffers from the same problems as the feature detector model criticized in the previous chapter. One of the main difficulties in vector coding is the enormous complexity of microfeatures that may be relevant for triggering any given concept. Macro-phenomenological identification and analysis of the task domain, based on folk psychology, are indispensable here. In nativist nets, the obvious procedure is first to decompose distal properties into microfeatures, then to use them for designing the weight matrix’s prototype vector, and finally to test the model for empirical adequacy. The analysis-design-test cycle may then be repeated in a continuous process of abductive refinement, using test results to correct the original analysis and design. In empiricist nets, the procedure is reversed, but consists of the same steps: test results are analyzed in terms of design, guided by phenomenological task analysis. The model may then be adjusted in virtue of these top-down considerations.

Although top-down analysis is a necessary condition for designing and understanding network connectivity, it is also *insufficient*. When considered in isolation, the top-down approach suffers from the same problem as its bottom-up counterpart, namely, *lack of constraints* on the determination of microfeatural relevance. As we saw in the previous chapter, many different mechanisms may be equally capable of subserving a given cognitive function. In the present context, this translates as the problem that many differ-
ent sets of microfeatures may be constitutive of a given concept (prototype vector). Isolated top-down analysis has no access to natural constraints posed by considerations of organizational, computational and implementational feasibility; these factors show up only in bottom-up analysis. Clearly, if the problem of selecting suitable microfeatures can be solved at all, it will be by availing oneself of all the information pertinent to the subject. A rich source of inspiration may be the brain, which has already solved the problem for its own purposes, so to speak. Hence, study of the brain may supply important hints and corrections for refining folk psychology’s abstract picture of cognition. Here lies the main attraction of connectionism: that it can integrate high-level task analysis and low-level understanding of neural wetware.

This general aspect of connectionist modeling is highlighted by the last three points mentioned above, (24)-(26). Together, they stress the fact that neural nets have two faces: one natural, the other hermeneutical. These faces correspond to the two aspects of the propositional attitudes discussed in chapter three. In order to be cognitively relevant, connectionism is required both to respect the epistemic aspect of propositional attitudes (meaningful interaction with environment), and to elaborate upon their natural aspect (description of internal states).

The requirement of conceptual interaction with folk psychology may take the form of implementation as well as that of correction. Some caution is called for. Referring back to an argument developed in the previous chapter, we should not be blind to the methodological risks involved in taking folk psychology at face value. As an example I mention the model for conceptual redeployment. It may well be that conceptual revolutions occurring primarily at the social and linguistic level have no immediate counterpart at the level of individuals, although folk psychology would tend to make us believe otherwise. Certain aspects may be better explained by sociology of knowledge than by connectionism. In cases like these, it would be wrong to simply project the relevant social mechanisms onto the neurophysiology of individual human beings. Notice, though, that the same objection applies to classical cognitive psychology. On the other hand, connectionism can make important contributions to the correction of folk psychology. Consider the example of perceptual recognition, categorization, and explanatory understanding, discussed above. Three phenomena which counted as dissimilar by the standards of folk psychology are computationally reduced to a single model of prototype activation. Yet, it would be absurd to maintain that explanation has been eliminated by
connectionism; rather, the original notion of explanation has been vindicated, corrected, and refined. The upshot of this example is that distinctions in folk psychology should not be carelessly projected onto neural nets; it does not follow, however, that folk psychology is to be wholesale mistrusted.

I close this section with raising a possible objection to the view defended here. It may be argued that we should distinguish between, on the one hand, the global, functional organization of the mind as described by folk psychology, and, on the other hand, the specific constraints posed on the semantics of mental symbols. Although connectionism is indeed able (and required) to adopt folk psychology’s functional inventory of mental explananda, as I have argued, it may still fail to account for the semantics of mental symbols. This is the subject of the final section.

4. Connectionism and mental content

Critics of connectionism have argued that network architecture is framed at the wrong level of analysis to account for the semantics of mental symbols. In this respect, connectionism is claimed to be incommensurable with the concerns of folk psychology and ‘classical’ cognitive science. Criticism centers on the absence of explicit rules and articulate representations in connectionist models, a fact that is often stressed by advocates and critics alike. Particularly influential in this context is an argument advanced by Fodor and Pylyshyn (Fodor 1987, 135ff; Fodor and Pylyshyn 1988). I think that the debate on the semantical status of connectionism reveals important aspects of cognitive science generally, irrespective of one’s choice of architecture. In this section I argue that the semantical controversy instigated by Fodor and Pylyshyn hinges on a false dilemma between ‘mere implementation’ and ‘downright elimination’, a dilemma that is ultimately based on a presumption that mental content must be intrinsic. The notion of intrinsic content will be scrutinized further in subsequent chapters.

*The constituent structure of thought*

Fodor and Pylyshyn start their argument from the observation that an adequate understanding of cognitive performance, as specified by folk psychology, is possible only if we assume that mental representations are internally structured, and that mental processes are sensitive to this structure. More specifically, they submit that mental representations require a combinatorial syntax and semantics, analogous to that of conventional human
languages. There are three reasons for this claim (Fodor and Pylyshyn 1988, 33ff). First, cognitive competence is typically productive: we can understand and produce a potentially infinite number of different propositions. Since this capacity is achieved using only finite resources, it must rely on recursive operations requiring a combinatorial syntax of thought. Secondly, cognitive competence is typically systematic: there are systematic connections between our capacity to comprehend one thing and our capacity to comprehend others. For example, anyone who can think ‘John loves the girl’ can also think ‘The girl loves John’. This means that “the two representations, like the two sentences, must be made of the same parts” (op. cit., 39). Finally, cognition is typically coherent from a syntactic and semantical point of view. For example, if you believe that turtles are slower than rabbits, and that rabbits are slower than Ferraris, you are typically prepared to infer that turtles are slower than Ferraris. Similarly, from a true conjunction ‘A&B’ you readily infer that both conjuncts ‘A’ and ‘B’ are true. Again, coherence requires that mental representations have compositional structure, and that mental processes be sensitive to this structure.

Fodor and Pylyshyn conclude from these considerations that cognitive science is committed to the existence of a ‘language of thought’, a representational medium that is roughly equivalent to conventional (formal or natural) languages (cf. Fodor 1975 and 1987, 135ff). The language of thought is supposed to consist of atomic symbols from which composite ones are formed, and which can be transformed into other symbols in accordance

\[
\begin{array}{c}
\text{1} \\
\text{[A&B]} \\
\text{2} \\
\text{[A]} \\
\text{3} \\
\text{[B]}
\end{array}
\]

*Figure 4.14: Connectionist inference generator according to Fodor and Pylyshyn*

As indicated by the labels on the nodes (between square brackets), the operation of this net may be interpreted as d
with the grammar of mind. The semantics of this combinatorial language maps isomorphously onto its syntax. As Daniel Dennett (1981a) graphically put it, the syntactic engine of the mind drives a semantic engine. Now, according to Fodor and Pylyshyn, connectionism has no room for a language of thought, because the syntax and semantics of connectionist models are not combinatorial.

“Connectionist theories acknowledge only causal connectedness as a primitive relation among nodes; when you know how activation and inhibition flow among them, you know everything there is to know about how the nodes in a network are related. By contrast, Classical theories acknowledge not only causal relations among the semantically evaluable objects they posit, but also a range of structural relations, of which constituency is paradigmatic” (Fodor and Pylyshyn 1988, 12).

To illustrate this claim, the authors contrast a ‘classical machine’ for drawing inferences from ‘A&B’ to ‘A’ or ‘B’ (for example, a Turing machine) with a connectionist machine executing the same function. Figure 4.14 shows the connectionist machine referred to in the following passage:

“In the Classical machine, the objects to which the content A&B is ascribed (viz., tokens of the expression ‘A&B’) literally contain, as proper parts, objects to which the content A is ascribed (viz., tokens of the expression ‘A’). Moreover, the semantics (e.g., the satisfaction conditions) of the expression ‘A&B’ is determined in a uniform way by the semantics of its constituents. By contrast, in the Connectionist machine none of this is true; the object to which the content A&B is ascribed (viz., node 1) is causally connected to the object to which the content A is ascribed (viz., node 2); but there is no structural (e.g., part/whole) relation that holds between them. In short, it is characteristic of Classical systems, but not of Connectionist systems, to exploit arrays of symbols some of which are atomic (e.g., expressions like ‘A’) but indefinitely many of which have other symbols as syntactic and semantic parts (e.g., expressions like ‘A&B’)” (op. cit., 16).

The upshot of this argument is that the causal relations holding between nodes 1, 2 and 3 are insufficient to bring out the logical relations holding between the representations involved. The nodes as such have no constituent structure; hence, they do not explain inference. Still, they may be inter-
preted in terms of internally structured symbols. Seen from this perspective, the connectionist machine implements a virtual machine of classical signature, defined over articulate, syntactically structured symbols and explicit rules for compounding and transforming symbols. Unlike the bare connectionist graph of figure 4.14, the virtual machine brings out the cognitively salient aspects of the function computed. Generally speaking, connectionist architecture is said to be located at a level below that of symbolic processing. It describes a specific hardware implementation, but not the cognitive function computed by it.

Fodor and Pylyshyn conclude that connectionism is faced with a fatal dilemma. If network models are proposed as an alternative to classical theories, they are badly inadequate: far from explaining cognition, they may sooner be seen as an attempt to eliminate it. On the other hand, to the extent that network models are adequate, they are not really an alternative to classical theories: connectionist models may implement classical machines for a specific kind of parallel hardware, but they cannot replace them.

Rules and representations

Part of the force of the semantical objection derives from the fact that it rests on premises that are accepted by critics and advocates alike. Advocates of connectionism often make much of the claim that network models have no need for either explicit rules or explicit data structures. A trained-up network sets its own synaptic weights and transfer functions, adapting itself to the solution space of the problem. A classical machine, by contrast, has a built-in solution space defined by its program. The programmer has laboriously specified all the relevant rules and data to be used by the machine, based on a prior analysis of the problem to be computed. By this token, the classical machine has cognition written all over it, whereas the acquired structure of empiricist nets remains relatively opaque. (Notice, however, that the innate connectivity structure of network models is typically based on a prior understanding of the problem domain.)

From a practical point of view, the two kinds of machines exemplify two different styles of programming, each with its own advantages and drawbacks. The classical programmer is concerned with data stacks and rule hierarchies, with memory resources, loops and halting. His goal is to ensure that the instruction code and the data will guide the machine to the desired solution. The connectionist, in contrast, takes decisions about the number of layers and units to be used, about how to encode the inputs, about setting
the thresholds and the connection weights, and about which learning algorithm to apply. He then sits back and waits for the machine to settle in a stable configuration for him to analyze. In the classical case, it is the programmer who teaches the machine; in the connectionist case, it is the machine that teaches the programmer.

The practical differences between classical and connectionist programming are vast, but they do not entail that network models operate without rules and representations. The rules and representations in empiricist nets are just not built in. Rather, they are discovered as ‘emergent’ properties (cf. Hofstadter 1985; Smolensky 1987 and 1988; Cummins 1989, ch. 11). Earlier I pressed the need for connectionists to make explicit this emergent structure, that is, to analyze the systematicity in the net’s performance. By interpreting network activity it is discovered that groups of nodes act as distributed representations of salient features of the problem domain, that layer-to-layer transformations compute specific cognitive functions, that connectivity structure defines prototypes, and so on. As described above, there are various hermeneutical techniques available now (microfeature analysis, analysis of state and weight space partitioning, Hinton diagrams). The hermeneutical effort is absolutely essential for understanding network activity in cognitive terms: without the emergent structure, the net is just a bundle of meaningless vectors.

Now, from a computational point of view, it may be thought that emergent rules and representations are just not good enough. A computational account of cognitive competence needs machines that compute representations using rules at the same level as the representations themselves, so to speak. But computation in connectionist machines takes place below the representational level. Nodes compute local state transitions, not emergent symbols; by the same token, they compute by means of local transfer functions, not by means of emergent rules. Hence, referring back to Fodor and Pylyshyn’s argument, it may be thought that, whatever constituent structure the emergent syntax of network models displays, it is just not the object of connectionist computation. Therefore, the computation does not explain cognitive capacity. This objection is based on a misunderstanding, however. The elementary processes in network models operate on single nodes, but this does not mean that computation is not defined over distributed groups of nodes, over ensembles of layers, and even over nets as a whole. The cognitive allure of connectionist modeling, as explained in this chapter, resides precisely in the fact that it makes the interlevel relations
computationally transparent, so to speak. If local activation is computed, then so are patterns of distributed activation (vectors); similarly, if weights define local functions, then ensembles of weights (matrices) define global functions. That is what vector notation is for.

Mere implementation?

On the face of it, the above considerations seem to give Fodor and Pylyshyn exactly what they want: connectionist models implementing classical machines. It is the emergent machine, not the network as such, that cuts cognition at its joints. Considering the fact that the emergent machine is a virtual machine of classical signature, it would seem that connectionist architecture is indeed not novel: it is merely an implementational theory for classical machines running on brain-like hardware.

I think that this conclusion misjudges the issue, however. It rests on a false dilemma between networks being either irrelevant or merely implementational. According to Fodor and Pylyshyn, there are only two options open to connectionists: changing the subject (the eliminativist reading of connectionism), or doing cognitive science the classical way (the implementationalist reading). But in point of fact there is a whole range of alternative options available, which are all relevantly different from either pure implementationism or pure eliminativism. As intimated earlier (chapters 2 and 3), these alternatives may broadly be termed ‘revisionist’ approaches in cognitive science. Their aim is to combine the best of connectionism and classical symbol processing, opting for a continuous interaction between the conceptual resources of top-down research and bottom-up constraints. Examples of such revisionist approaches include ‘approximationism’ (Smolensky 1987 and 1988), ‘compatibilism’ (Touretzky and Hinton 1988), ‘externalism’ (Rumelhart, Smolensky et al. 1986; Smolensky 1988), the ‘limited cognitivism’ advocated by Clark (1989, chs. 7ff), and the attenuated form of reductionism recently espoused by Churchland (more on which in chapter 5).

I shall not go into a discussion of the various forms of revisionism here. (For a critical survey, see Bechtel and Abrahamsen 1991, 226ff.) Instead, I want to draw attention to their common insight, namely, that classical and connectionist constraints (or some similar ones) are equally indispensable. It would be a grave mistake to think that ‘pure’ connectionist architecture, without interpretation in terms of classical constraints, is able to explain cognition. But it is equally misguided to think that only ‘pure’ classical ar-
chitecture, cut off from all bottom-up constraints, is the legitimate concern of cognitive science.\textsuperscript{15} Implementation (‘mere implementation’) is not an achievement to be taken lightly. As I see it, connectionist machines do indeed implement classical machines, but not in a derogatory sense of the word. They would be mere implementations of classical machines only if we already had a complete theory of the machines to be implemented, that is, if cognitive structure were given to us a priori. This is obviously not the case. By contrast, if cognitive structure is to be discovered, then the vector approach of neural epistemics seems ideally suited for the task.

\textit{Intrinsic content and the labeling fallacy}

This brings me to a final aspect of Fodor and Pylyshyn’s argument. Classical cognitivists typically take it for granted that the abstract virtual machine is the real engine of cognition. This claim is usually backed up by a standard functionalist argument: the universal engine is set up in a realm sui generis, distinct from its various hardware implementations. Although I fully accept functionalism as a framework for co-ordinating research at different levels of analysis, I think it is wrong to use it as a dismissal of one research level in favor of another. Such a dismissal, for example in favor of a purely classical architecture, would imply that cognitive structure is ‘given’ to us independent of all hardware constraints. Yet, we have no such ‘direct access’ to the essence of cognition. The tendency to think otherwise strikes me as a pernicious prejudice in cognitive science. It is arguably a relic of the Cartesian metaphysics of subjectivity, according to which the essence of mind is immediately known to itself. In the context of semantics, this claim takes the form of a presumption that mental content is ‘given’ to the mind, or in more modern terms, that it is determined exclusively by factors intrinsic to the computational system. It is this presumption, I believe, that lies at the root of Fodor and Pylyshyn’s argument against connectionism.

To illustrate this point, let me return once more to the practical differences between classical and connectionist styles of programming. The operation of classical machines is determined by the ready-made structure built in by the programmer. If the programmer wants the machine to represent cats as being furry, this is typically accomplished by including in its program some relevantly structured string of symbols, say, ‘FURRY(CAT)’ or ‘IF CAT (THING) THEN FURRY(THING)’. By virtue of this style of pro-
programming, the structure of the classical machine appears as semantically and cognitively transparent. Connectionist structure, in contrast, tends to be more opaque—it is acquired rather than pre-programmed. As pointed out above, the difference between the two approaches is one between ‘explicit’ and ‘emergent’ structure. Still, there is cognitive structure in either case: once network activity is interpreted by the researcher, the relevant structure becomes just as explicit as it is in classical machines. In a typical network model, the intended interpretation is indicated by the labels on the nodes. For example, the three nodes in figure 4.14 are labeled ‘A’, ‘B’, and ‘A&B’, respectively. These expressions are obviously related in terms of constituency and entailment. Hence, connectionists can legitimately claim that network models are cognitively relevant, inasmuch as the cognitive significance of the net’s computational activity is revealed by the labels on the nodes.

This suggestion is rejected by Fodor and Pylyshyn, however. Although they admit that the expressions on the labels have the required constituent structure, they point out that the labels are not part of connectionist architecture itself.

"Strictly speaking, the labels play no role at all in determining the operation of a Connectionist machine; in particular, the operation of the machine is unaffected by the syntactic and semantic relations that hold among the expressions that are used as labels. To put this another way, the node labels in a Connectionist machine are not part of the causal structure of the machine" (op. cit., 17).

The mistake purportedly made by connectionists here may be called the ‘labeling fallacy’, that is, the conflation of node structure and label structure. According to Fodor and Pylyshyn, labels cannot save network models, because the network can read only the output of individual nodes but not the labels they carry. Change the labels, and the machine will continue to operate as before. Classical machines, by contrast, are said to be able to actually read the symbols they are given: different symbols make the system behave in different ways. Feed the classical machine ‘BALD(CAT)’ instead of ‘FURRY (CAT)’, and it will change its mind on cats. In more stately terms,

“the state transitions of Classical machines are causally determined by the structure—including the constituent structure—of the symbol arrays
that the machines transform: change the symbols and the system behaves quite differently” (ibid.).

The objection raised by Fodor and Pylyshyn reveals a deeply rooted presumption in cognitive science, namely, that cognitive research should be conducted from the internal point of view of the operator handling the symbols. Only on this assumption does it make sense to claim that the labels in connectionist models “are not part of the causal structure of the machine”. Apparently, Fodor and Pylyshyn require that the connectionist operator should be able to read the labels on the nodes, and then to act in accordance with the instructions he finds on them: the content of the symbols to be computed must be given to the machine itself if its operations are to qualify as a explanation of cognitive performance. On this condition, the ‘emergent content’ as displayed on the labels is not good enough, because it is not intrinsic to the computational system as such. Once interpreted by the researcher emergent structure may be explicit, but it is not explicit-to-the-machine.

I shall have more to say on the requirement of intrinsic content in subsequent chapters. Suffice it here to point out that it is too strong for any cognitive machine, connectionist or classical. Whether a classical machine is able to discriminate between ‘FURRY(CAT)’ and ‘BALD(CAT)’ tells us nothing about the cognitive function it is computing. ‘CAT’ may as well refer to dogs, and ‘FURRY’ may stand for barking—the ‘internal operator’ cannot tell the difference. Similarly, the connectionist operator is given only activation states and vectors to work on—he is ignorant of their epistemic function and meaning. In this respect, classical machines and connectionist machines are in exactly the same plight. Moreover, it is true that the change from ‘FURRY (CAT)’ to ‘BALD(CAT)’ affects the performance of classical machines, as Fodor and Pylyshyn correctly point out. But, again, the same is true of connectionist machines: change a net’s connectivity structure and it will perform very differently, requiring a readjustment of the labels on the nodes.

None of these remarks should be construed as claiming that connectionist architecture is not novel. My point is rather that adopting a connectionist framework does not force us to abandon the ‘classical’ idea that cognition is the computational manipulation of semantically structured symbols. If this requires us to give up the idea that mental symbols must be ‘given’ to the internal operator, I take this to be progress.
Chapter three
Eliminative materialism and folk psychology

One of the most mind-boggling positions in modern philosophy is doubtless that of eliminative materialism. Its claim is as simple as it is bold, namely, that we do not have minds. We are, in reality, not endowed with anything like the beliefs and desires that are commonly held to drive our behavior. For scientific purposes at least, our talk of mind should be eliminated. More technically put, eliminative materialism makes the following claim,

“that our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience” (Churchland 1981, 67).

Over the past fifteen years, this position was developed by Paul Churchland, who is still its most ardent defender. The thesis made its first explicit, published appearance in Churchland’s book, *Scientific realism and the plasticity of mind* (1979), soon followed by a famous paper in the *Journal of Philosophy* (Churchland 1981). Earlier suggestions in the same direction had been made by Paul Feyerabend (1962; 1963a; 1963b) and Richard Rorty (1965; 1970). More recently, philosophers such as Daniel Dennett (1978; 1987; 1991) and Stephen Stich (1983) have defended positions that bear unmistakably eliminativist traits, Dennett’s version being close to instrumentalism about the mental, while Stich’s is syntactic as opposed to Churchland’s neuroscientific brand. I shall concentrate here on Churchland’s version, which is by far the most articulate and straightforward defense available.

In this chapter I am concerned primarily with the *eliminative* part of eliminative materialism. Other aspects of Churchland’s philosophy, most notably his materialism, his naturalism, and its connection with connectionism, will be reserved for discussion in later chapters. Regarding Churchland’s eliminativism, I shall be considering the following three questions.1
1. Are there reasons for doubting the validity of our common-sense conception of mental phenomena? In other words, is it plausible that folk psychology is false?

2. Is it prudent to endorse a research strategy in cognitive science that systematically ignores folk psychology?

3. Is the ‘bottom-up’ approach of neuroscience, advocated by eliminative materialism, empirically successful at explaining cognitive phenomena?

Each of these questions tracks a different aspect of eliminative materialism’s place in the constellation of contemporary science and philosophy. With a growing number of kindred spirits, I hold that no philosophy, and in particular no theory in the philosophy of science, can be adequately evaluated without due consideration of the scientific evidence it calls upon, or the scientific theories it advocates. This also applies to eliminative materialism, which is specifically aimed at defending a certain research program in cognitive science. In evaluating Churchland’s position, I avail myself of the descriptive framework for analyzing a theory’s merits and weaknesses as developed by Lakatos (1970; 1978), Laudan (1977), and others. In their terminology, the questions raised above address the metaphysical (or conceptual), methodological, and empirical plausibility of eliminative materialism.  

The argument to be developed in this chapter is fairly straightforward. First, the main arguments on behalf of eliminative materialism are set forth (section 1). In the sections that follow, I discuss the conceptual, methodological, and empirical credentials of eliminativism, as indicated. Within this purview, I argue that folk psychology is relatively ‘observational’ and theoretically innocuous in character (section 2), that its descriptive vocabulary is methodologically indispensable (section 3), and that, without its help, cognitive science is headed for empirical infertility and stagnation (section 4). Finally, I shall try to tie these lines together in what may be called a ‘dual aspect’ view of folk psychology (section 5). The dualism proposed here is not ontological but epistemic in character. Subsequent chapters will trace the ramifications of this view in various other discussions in contemporary philosophy of mind.

1. Eliminative materialism

The fallibility of folk theory

In traditional philosophical analysis, the idea that our everyday claims about the mind are false would be simply preposterous. Consider a report of your
own state of mind, such as regretting you ever opened this book. The claim
that this is how you feel about my book is generally held to be beyond falsi-
ification, partly because it is based on direct observation, partly because it is
concerned with a private and incorrigible episode of your inner life. In order
to appreciate eliminative materialism, we must leave this traditional analysis
behind, and make a giant leap into twentieth-century philosophy of knowl-
edge. I shall briefly sketch some of the philosophical background from which
eliminative materialism has sprung.  

A common theme in the theories of knowledge developed by Quine (1953,
1990), Sellars (1963), Feyerabend (1962), and others is the claim that all of our
knowledge is thoroughly theoretical and fallible. Each of our beliefs is a fallible
hypothesis about reality. In framing our thoughts, we use concepts that may
or may not be appropriate. Some sets of concepts perform better than others,
which they may then replace. Our conceptual systems are constantly being
tested and revised; there is never a set of concepts that cannot be improved
upon. In this sense, each belief can be revised; hence, all beliefs are hypotheti-
cal and fallible. This applies even to those beliefs that are vested in our so-
called ‘observation statements’. No matter how spontaneous, non-inferential or
intuitively evident they are, even our observational beliefs remain conjectures
that can in due course come to be revised. The conceptual component that is
incorporated in all our beliefs cannot be reduced to anything ‘given in direct
experience’. Common-sense beliefs are no exception to this; in the above sense
of the word they are as ‘theoretical’ as scientific hypotheses. Our common-
sense view of the mental, for example, makes use of an elaborate system of
concepts such as ‘belief’, ‘desire’, ‘expectation’, ‘hope’, ‘fear’, etc. In order to
describe, predict and explain the behavior of ourselves and our fellow human
beings, our common-sense psychology postulates the existence of internal states
such as beliefs and desires. With a term of Bertrand Russell (1940), these inter-
nal states are called propositional attitudes. Now, when in everyday life we want
to explain why a person acts the way he does, we typically attribute to him
(sets of) specific propositional attributes, which we then (mostly implicitly)
subsume under some general rule. For example, if we want to understand
why Mary goes on a diet, we attribute to her a belief that she is overweight, a
desire to be slim and attractive, a belief that going on a diet is the right way to
achieve this, and so on. Moreover, we assume (general rule) that persons with
these internal states tend to act the way Mary does. In general, our common-
sense psychological explanations follow a nomological-deductive pattern,
based on a web of interrelated, lawlike generalizations of the following sort:
4. \((\forall x)(\forall p)(\forall q)\) \{(x \text{ hopes that } p) \& (x \text{ believes that } (p \text{ then } \neg q)) \& \text{ normal circumstances } \rightarrow (x \text{ hopes that } \neg q)\}

5. \((\forall x)(\forall p)(\forall q)\) \{(x \text{ believes that } p) \& (x \text{ believes that } (p \text{ then } q)) \& \text{ normal circumstances } \rightarrow (x \text{ believes that } q)\}

6. \((\forall x)(\forall p)(\forall q)\) \{(x \text{ desires that } p) \& (x \text{ sees that } \neg p) \& \text{ normal circumstances } \rightarrow (x \text{ is disappointed to find that } \neg p)\}\)

These generalizations are fallible empirical hypotheses. Moreover, the concepts they employ are defined in part by their place in the overall system of laws. Following usage, the ensemble of lawlike generalizations, rules of thumb and interconnected concepts I refer to as ‘folk psychology’.

Folk psychology forms part of a much larger corpus of common-sense beliefs about ourselves and the world, which express themselves in our everyday talk about various aspects of reality. By analogy to folk psychology, I shall refer to these sets of beliefs as ‘folk theories’.

These remarks suffice to clear the ground for eliminative materialism. I conclude here with some final epistemological preliminaries to remind the reader where the onus of proof lies. Though folk theories are fallible, this does not imply that they are false. Until specific reasons for doubting their validity are given, the possibility of their being false remains merely abstract. Global skepticism of the Cartesian kind would result in acute epistemic paralysis; contrary to what Descartes led us to believe, doubting all our beliefs at once leaves us without a starting-point for believing anything at all. As Wilfrid Sellars once remarked,

“empirical knowledge, like its sophisticated extension, science, is rational, not because it has a foundation, but because it is a self-correcting enterprise which can put any claim in jeopardy, though not all at once” (Sellars 1963, 170).

Doubt is always local. It uses certain beliefs as an Archimedean point for casting doubt on others (cf. Rescher 1980). This is precisely what eliminative materialism intends to do. Taking modern natural science as its starting-point, it seeks to give specific reasons for doubting, from this purview, the validity of folk psychology. Let us now turn to these specific reasons.


Arguments for eliminative materialism

Why should we believe that what folk psychology has to offer is a “false and radically misleading conception of the causes of human behavior and the nature of cognitive activity” (Churchland 1988a, 43)? Churchland offers various arguments in defense of this claim. They fall in essentially two categories: those that are concerned with refuting defenses of folk psychology and traditional cognitive psychology, and those that challenge folk psychology itself. Since the onus of proof lies clearly with eliminative materialism, the first class of arguments can be ignored for the moment. An example here is Churchland’s refutation of functionalism, which will be discussed later. The arguments of the second category, which contain the real challenge to folk psychology itself, can be summarized as follows.  

7. Most folk theories have proved false; therefore, it is unlikely that folk psychology will turn out to be true.
8. Folk psychology is an empirically and conceptually degenerating research program; as such, it deserves to be terminated.
9. There is a vastly superior competitor to folk psychology, namely, the new research program in cognitive neuroscience.

The first argument is based on an inductive generalization. According to Churchland, most of our folk theories proved so vague and primitive that, in due course, they were replaced by scientific theories of an altogether different nature. Differences between the old and new theories were so profound that the venerable ontologies of folk theory have been eliminated rather than smoothly incorporated into the new ontology of science. Typical examples include primitive explanations of the animate and the inanimate, theories of the animal and vegetable world, hypotheses concerning demons, nymphs, satyrs and witches, ancient explanations of celestial mechanics — they all turned out to be false and fundamentally misguided. Not only have the theories been discarded, but also the ontology on which they relied. If this is the fate that befalls folk theories generally, why should we expect our prescientific notion of mind to fare any better? “It would be a miracle if we had got that one right the very first time, when we fell down so hard on all the others” (Churchland 1988a, 46).

It is interesting to note that Churchland’s attitude towards folk theory bears a certain resemblance to the ‘three stage’ doctrine of Auguste Comte, the 19th-century father of positivism, according to whom all thought, in the course of
its evolvement, must pass through a theological and a metaphysical stage before reaching the truly scientific, positive stage (Comte 1830). The objection to folk psychology, in this context, would be that it is somehow tied to the metaphysical stage. Though no longer depending on supernatural agencies to explain the phenomena of cognition (as the theological stage would have it), folk psychology has replaced them by abstract ones, reifying the concepts and principles found in ordinary language, that is, the propositional attitudes. Now, Churchland seems to be urging that we should turn to the positive stage at last, eschewing the search for inner natures and essential causes. This positive stage is that of mature neuroscience.

The comparison between Churchland and Comte may go a long way, yet it is also somewhat misleading, suggesting as it does that folk theories are mere metaphysical superstition. Should folk psychology be eliminated, it finds itself in the company of numerous, perfectly respectable, scientific hypotheses. Should mind be eliminated, it shares the fate of many other theoretical entities, including the caloric or heat fluid, Cartesian vortices, and Newtonian and Maxwellian aether (see, for example, Bynum et al., 1981). It is, in sum, by no means inconceivable that folk psychology is false, but this does not automatically mean that it was complete nonsense to start with.

Churchland’s second argument draws attention to specific deficiencies of folk psychology. Allegedly, its empirical and conceptual deficiencies are so serious that folk psychology is called a “stagnant or degenerating research program”, the history of which is marked by “retreat, infertility and decadence” (Churchland 1981, 74-75; cf. Lakatos 1970 and 1978). On the one hand, folk psychology shows a notorious lack of empirical progress. Whatever empirical success has been accomplished over time is due to a set of standard explanations of a relatively small range of phenomena. Apart from these, no novel facts have been predicted, nor have any new predictions been confirmed. At the same time, folk psychology is beset with a large variety of anomalies, none of which have been adequately answered, in spite of its very long career. Even such basic mental capacities as memory, learning, and intelligence, not to mention mental diseases, creativity, or sensorimotoric coordination, are still as much of a puzzle today as they were in ancient Greece. No scientific theory with a two thousand year record as bad as folk psychology’s would have been treated so indulgently. Rather, it would have been eliminated.

Apart from these empirical difficulties, folk psychology also suffers from serious conceptual problems. Its use of ‘intentional’ categories is entirely out of tune with the rest of our physicalist science. We now have the outlines of a
methodologically and ontologically homogeneous world view, spanning from evolutionary biology, neuroscience, and biochemistry, to quantum physics and astronomy. In this scheme of things there is also room for *Homo sapiens*, as studied from a physicalist perspective. But folk psychology “is no part of this growing synthesis. Its intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus” (Churchland 1981, 75). The ‘intentionalism’ in cognitive science, Churchland argues, should share the fate of ‘vitalism’ in biology, and land on the conceptual junkyard of discarded theories.

This brings us to the third and final argument, summarized in claim (9) above. According to Churchland, there is a viable, radical alternative to folk psychology based cognitive psychology, in the form of a research program in neuroscience that is variously called ‘parallel distributed processing’ (Rumelhart and McClelland 1986a), ‘connectionism’ (Smolensky 1988), or ‘natural computation’ (Richards 1988). Churchland argues that this program’s proven fertility warrants great expectations for the future. Its progress is in fact so steep, that already it outshadows its stagnant competitors, and folk psychology in particular. Folk psychology’s use of intentional categories prevents it from being able to tie in with the rest of physicalist science, and with neuroscience in particular. On account of this, intentional psychology cannot profit from the new results in connectionism. On the contrary, it stands in grave danger of being eliminated by it.

Classical cognitive psychology and modern connectionist neuroscience are generally taken to be incompatible frameworks, perhaps even to the extent that they are evidentially and conceptually *incommensurable*. Rather than being worried by this aspect of cognitive neuroscience, Churchland hails it as a benefit of the approach. If the programs seem to be working at cross-purposes, this is all for the best. For the intentional idiom connectionism lacks is the probable cause of folk psychology’s degeneration, while its absence in connectionism may explain the latter’s success. Instead of trying to explain the familiar intentional phenomena, connectionism had better define its own explananda; instead of trying to meet folk psychology’s requirements, it had better define its own standards of explanatory success. If this analysis is correct, the prospects are that folk psychology will simply *disappear* from the scientific stage, although it may continue to be used for everyday purposes.
2. Conceptual or metaphysical arguments

The more things change, the more they stay the same

The implications of eliminative materialism for cognitive science and our view of the mental have slowly come to be appreciated over the past ten years. In this section, I take a closer look at the arguments summarized above. First, I address Churchland’s more general claims with regard to folk theories, as presupposed by his first and second argument. In response to these, I sketch an alternative account of the relation between folk theory and science, arguing that eliminative materialism misses out on some essential differences between the two. Finally, I focus more specifically on the charge of folk psychology’s intentional categories being incompatible with physicalism (second and third argument).

Churchland’s first argument, an inference from folk theory to folk psychology, seems to be based on a false generalization. It is simply not true that the vast majority of our past folk conceptions have been eliminated rather than reduced by scientific theories. Rather, folk theories have typically survived most scientific theories in their domain; the latter come and go, but folk views stay. The same goes for the vast majority of folk entities, as is obvious from countless examples. The various species of plants and animals, soil and snow, seas and clouds, bodily organs, minerals and almost all other things: old scientific speculations may have been replaced by new ones, yet the bulk of our common-sense views and taxonomies have remained intact, and have even been merged surprisingly smoothly with the new sets of theories. One or two insects move to a different family, and perhaps one or two stars become planets, but as far as folk theory is concerned nothing more drastic happens.

Churchland’s inference is based on a cardboard version of the history of folk theories. Once the image is corrected, the pessimistic prospects for folk psychology are much improved. By the same token, Churchland’s ‘miracle’ argument looses its validity. The question confronting us is not what is so good about folk psychology that it should be exempted from the global elimination of folk views. We should ask instead what is so bad about it that it should be relinquished.

The shortcomings of Churchland’s second argument are of a more fundamental nature. From the premise that folk theories are like scientific theories in being fallible empirical hypotheses about reality, Churchland infers that they are like scientific theories in all other respects as well. In particular, he infers that folk theories should meet the same standards of evaluation as scientific theories.
As we have seen, Churchland invokes Lakatos’s notion of the methodology of scientific research programs. From this purview, he argues that folk psychology is both empirically and conceptually degenerating, thereby assuming that it can be reduced or eliminated in the same way as scientific hypotheses. Yet, there is reason to resist this implication. In the first place, the argument is fallacious as stated. From being like X in one respect, it does not follow that something is like X in all respects. Churchland still owes us some positive reason for extending the analogy as far as he does. In addition, and more importantly, there are indications that folk theories are not like scientific research programs in the relevant respects. I shall try to develop two essential points of difference here. The following remarks do not pretend to be anything like a full-blooded account of the relation between common-sense and science. Yet, they raise some points that, in my opinion, any plausible account of this relation should address, and that are blatantly ignored by eliminative materialism.

In the first place, as compared to scientific theories, folk theories are marked by a high degree of observationality. Typically, their empirical generalizations are of a relatively shallow nature. In Quine’s famous image of knowledge as a field of force whose boundary conditions are experience, the pronouncements of folk theory are close to the boundary (Quine 1953; 1970; 1990). Although containing an irreducibly conceptual or ‘theoretical’ component, the concepts they employ are few and simple. Folk theoretical sentences tend to be the ones that are used in learning a language. Also, folk theoretical concepts tend to be the ones in which observation sentences are couched. This sets folk theory apart from science, yet without impugning its fundamental fallibility and revisability.

In the second place, the descriptions and explanations used in everyday life with regard to the various domains of reality tend to be highly elliptic, or abstract. They typically give a prima facie delineation of the phenomena to be explained in a given domain of reality: they identify the explananda, rather than giving the explanations themselves. Perhaps it is proper to say that folk theories give the outline of an explanation, but the actual explanantia are left undetermined. It is for science, philosophy, religion, or whatever, to decide how they should be filled in. In combination with the former point, this implies, inter alia, that folk theories provide the observational identification of the material objects of a scientific domain, those phenomena that science should explain. Of course, as science changes, the formal object of science may change as well, and the identification of explananda may be adjusted. None of the
above implies that folk theory should be immutably fixed. What does follow, however, is that whatever revisions may prove necessary must always be piecemeal. It is impossible for any science, in any domain of reality, to change the entire observational vocabulary at once, without robbing itself of its explananda and thereby changing the subject. Hence, folk theory cannot be wholesale eliminated.

As indicated by the above two points, the relation between a folk theory $F$ and a scientific theory $S$ about the same domain of reality is essentially different from that between two competing scientific theories $S_1$ and $S_2$ about the same domain. In this respect, folk theories tend to be much more folk than theory. In the case of $S_1$ and $S_2$, which are explicit, systematic and articulate theories, it is more or less easy to establish what their points of agreement and their points of difference are, and to which extent they complement, reduce, or eliminate one another. The relation between $F$ and $S$, however, is much less

![Figure 3.1: Different status of folk theories and scientific theories](image)

The status enjoyed by folk theories about a given domain of reality is different from that of scientific theories about the same domain, as well as from the general cultural background provided by philosophy, religion, and similarly global sets of belief. Elimination and reduction may occur between competing scientific theories $S_1$ and $S_2$, but relations between folk theories and scientific theories are much less straightforward; typically, they tend to take the form of continuous conceptual interaction and coexistence. This fact may be explained by the high degree of observationality of folk theory, which serves as a prima facie demarcation of explananda rather than giving explanations itself.
Eliminative materialism straightforward. It is not one of elimination or reduction, but rather one of permanent conceptual interaction, S using F-concepts for describing the relevant domain of reality, refining them and enlarging upon them by introducing new concepts and explanations, thus influencing the meaning of F’s original concepts, yet without ever quite losing touch with them.

Let me give a simple example. In the course of time, our folk biology somehow developed into the science of biology. Today, modern infracellular biology describes its phenomena in a vocabulary (including ‘cells’, ‘membranes’, ‘protoplasm’, ‘ribosomes’, ‘organellae’, not to mention ‘RNA’, etc.) that our distant ancestors would have been quite unable to understand. The conceptual resources of scientific biology have obviously diverged from those of folk biology, and yet it would be wrong to conclude that the old folk entities have been eliminated, or that modern biology could have developed without continually basing itself on the concepts of folk biology. It would be equally wrong to say that folk biology has been reduced, in the strict sense of specifying lawlike type-type correlations between folk biology and scientific biology. What happened is rather that the science was born from folk theory, using the latter’s concepts to define its explananda, and proceeded to couch its explanations in terms of a set of newly developed concepts, thereby affecting in its turn our folk theoretical notions. There is no elimination or reduction, but only conceptual interaction, which is what warrants our conviction that even the findings of molecular and infracellular biology are about the same living things we deal with in everyday life.

Applying these remarks to folk psychology, we find that its lack of depth and problem solving capacity is hardly surprising. It offers not so much a theory of the mental, but rather a first, relatively observational, inventory of what there is to be explained in the first place. This new perspective undermines some of the central claims of the eliminative materialist. I shall briefly review its consequences for Churchland’s arguments.

In the first place, the alternative account explains why most folk theories (pace Churchland) have survived subsequent revolutions in science. It is only to be expected that folk psychology will follow the same pattern. Whether the prevailing philosophy is functionalist, reductionist or behaviorist, and whether our science is cognitive psychology, connectionism or neuroscience, folk psychology will remain relatively untinged by any such developments, acting rather as a kind of shared point of reference for all parties involved. I hardly need to point out that this seems to capture precisely what is going on in cognitive science and philosophy today.
Furthermore, we find that the application of scientific standards of theory evaluation to folk theory is misguided. Folk psychology is a “multi-purpose tool”, as Stephen Stich has pointed out (1983, 212-214). It was designed for various purposes, none of them scientific, and it serves them well. The pressure for scientific improvement is negligible. Folk theories do not lag behind in predictive and explanatory fertility; they are simply not in the race. If folk psychology has been holding out the explananda of the mental for over 2000 years, and there is still no explanation, then this is the fault of the science that has all this time failed to address these issues.

Finally, folk psychology and cognitive science are largely continuous with one another. If this were not the case, cognitive science could not be said to address the familiar phenomena of cognition. Viewed from this perspective, eliminative materialism seems to defy its own purpose. On the one hand, it wants some future developments in neuroscience to explain our cognitive abilities. But on the other hand it refuses to accept our standard, observational vocabulary for specifying these abilities. So how could we ever decide on what there is to be explained in the first place? This quandary is exemplified in the discussion on the status of connectionism, the radical alternative to cognitive psychology that has recently become available, and of which Churchland is an ardent defender (Churchland 1989a; 1990; 1992a; 1992b). Critics typically argue that connectionist theories are unable to explain cognition, either because they fail to address the proper explananda, or because they deal with only one specific, brain-like implementation of cognition. An evaluation of this criticism will be the topic of chapters four and five. There I will argue that connectionism can and must make use of the descriptive resources of folk psychology in terms of which cognitive phenomena are specified. Moreover, particularly in his more recent work, as we shall see in chapter five, Churchland himself seems to subscribe to the kind of conceptual interaction between folk psychology and connectionism suggested above, which tends to make his position reductionist rather than eliminativist.

The view advocated here resembles the moderate form of epistemic foundationalism proposed by Barbara Von Eckardt in her criticism of Churchland.

“Common-sense foundationalism claims that folk psychology is inadequate only in scope and depth; that is, it is not inadequate in the sense of being mistaken in any sense. (...) Thus, according to common-sense foundationalism, insofar as folk psychology is correct as far as it goes, any adequate science of cognition will use folk psychology not only as a starting
point but as a continuing constraint on what will count as an adequate explanation” (Von Eckardt 1984, 74-75).

I fully agree with this position, provided the proper caveats and disclaimers are observed. As was pointed out above, the ‘foundation’ on which cognitive science rests is thoroughly fallible; similarly, the ‘continuing constraint’ is subject to permanent scrutiny, and can always be revised, though only piecemeal, as we have seen.

Incommensurability?

This leaves us with Churchland’s final argument, which is based on the superiority of the new connectionist research program, incommensurable with folk psychology’s intrinsic intentionalism. In sections 3 and 4, we shall see that there is some reason to be skeptical about the prospects Churchland claims for his new program. As a matter of fact, with regard to its conceptual and empirical fertility, cognitive neuroscience is arguably a better example of a stagnant research program than folk psychology. Yet, the fact remains that folk psychology’s intentionalism might well be incompatible with physicalism generally. What if the phenomena singled out as explananda by folk psychology are simply wrought from the wrong conceptual grid? How could any science ever hope to make sense of them? In many respects, this seems to be the true worry motivating Churchland’s crusade against folk psychology. The analogy with vitalism in biology, frequently invoked by Churchland, points in this direction. Persuaded by the idea that ‘life’ is somehow special, biologists in the past were led to believe in some form of ‘vital principle’, such as the Aristotelian and Scholastic psyche, Van Helmont’s archeus, Blumenbach’s nisus formationis, Bergson’s élan vital, or Hans Driesch’s entelechy (Bynum et al. 1981, q.v.). Vitalism cleaved a sharp boundary between the animate and the inanimate. From the vitalist’s point of view, no science that lacked the concept of vitality could ever hope to explain the phenomena of life. Its conceptual and evidential repertoires were simply incommensurable with those of vitalism. Still, with the rise of organic chemistry, the differences between animate and inanimate beings came to be explained in terms of an underlying, common structure. The awkward concept of vitality was eliminated, and natural phenomena were taxonomized in radically new ways (Churchland 1982; 1979, 109-110).

A first reply to this objection would be that intentional categories may be awkward from the point of view of natural science, yet they perfectly fit the social sciences (cf. Stich 1983, 212-214). Perhaps Churchland would welcome
the latter’s elimination as well, but that hardly improves his situation, as he
does not offer an alternative to traditional social science. Yet, I believe a more
principled answer can be given, drawing on the elliptical or abstract nature of
folk psychology introduced above.

*Intentionality*

Due to the lack of textbooks on the subject, it is surprisingly hard to agree on
a canonic version of folk psychology.⁸ Does folk psychology commit us to
something like intentionality? And if so, what is it? ‘Intentionality’ is notori-
ously a philosopher’s term, not a part of common-sense. To a first approxima-
tion, it is the property of mental states which makes them be ‘about’ some-
thing. Thus, in a recent encyclopedia on the mind we read,

> “Intentionality is *aboutness*. Some things are about other things: a belief can
> be about icebergs, but an iceberg is not about anything; an idea can be
> about the number 7, but the number 7 is not about anything; a book or a
> film can be about Paris, but Paris is not about anything” (Dennett and
> Haugeland 1987, 383).

Let me give a concrete example. Suppose our aim is to explain why Luke
Skywalker is anxious about fighting Darth Vader. We attribute to him, *inter
alia*, the belief that Darth Vader is his father, that Darth Vader is evil, that evil
should be fought, and that a son should not fight his father (all of which are
true). Given these internal states, familiar folk psychological mechanisms ac-
count for the anxiety overtly manifested in Luke’s behavior. Now, what *is* this
belief we attribute? It is taken to be an ‘internal state’ in which Luke finds
himself. This state is characterized as a propositional attitude: an epistemic
attitude toward a mental symbol of a certain propositional content, analyzed
as follows:

10. (Luke Skywalker) (believes) (that Darth Vader is his father),

or more generally,

11. (subject) (epistemic attitude) (propositional content).

The general schema shows how intentionality, in the sense of ‘aboutness’,
enters into the notion of propositional attitudes. The internal states attributed
to a subject are identified in part by their aboutness. Moreover, it is this aboutness, or propositional content, by which internal states are systematically related to one another.

Assuming the above analysis to be correct, what is there for Churchland to take exception to? On his interpretation of folk psychology, it commits us to an abstruse metaphysical principle of intentionality comparable to the vital principle of biology, thus cleaving an insurmountable gap between cognitive phenomena and the rest of nature. Yet, it would seem that this analogy is not quite fair. Should not folk psychology be compared to folk biology, rather than to the science of biology? If this is so, the situation appears to be quite different from that pictured by eliminative materialism. Although it is true that biologists in the past, up into the nineteenth century, posited a metaphysical principle of vitality, folk biology is free from this error. If in everyday life we speak of things as ‘being alive’ or as ‘having life inside of them’, this is hardly because we are doing metaphysics. All we do, in fact, is to distinguish between two kinds of phenomena, those involving living things and those involving lifeless things. The explanation of the difference between the two is left to philosophers and scientists. Analogously, in the case of folk psychology, a distinction is made between cognitive and non-cognitive phenomena. What there is inside our heads that could explain the difference is left entirely open. The answer to that question is delegated to philosophers and scientists, who have variously suggested that it is an immaterial soul, a non-physical principle of intentionality, or a highly complex nervous system. Of this indulgence, however, folk psychology appears to be innocent.

The answer given here follows my earlier analysis of folk psychology as being largely theoretically innocuous. I propose the following, minimal definition of intentionality. To a first approximation, for any system S,

12. \( S \) has intentionality = \( \text{def} \) \( S \) is capable of systematic interaction with its environment, such as is best captured by the assumption that its behavior is determined by internal representations of the environment.

This minimalist definition is obviously close to Dennett’s notion of an ‘intentional system’, that is, a system toward which it makes sense to take an ‘intentional stance’ (Dennett 1978; 1987). The nature of the representations in definition (12) is left open; they may be neural, connectionist, iconic, propositional, functional, irreducibly mental, or whatever. Of course, it remains to be seen whether the notion of a ‘systematic interaction’ can be independently specified,
and whether the notion of ‘representation’ can be naturalized (on which more in later chapters) — but these are separate issues, not to be confused with that of the status of folk psychology as such.

Churchland seems to have no quarrel with the notion of internal representation as such. He would probably welcome definition (12), adding to it a *neural* interpretation of the nature of the representations. However, he thinks the concept of intentionality employed by folk psychology is significantly *richer* than suggested here. In addition to the analogy with vitality, he argues that intentionality is wed to the notion of *linguistic* representation. It is a capacity for processing internal representations in a *language of thought*. As indicated above, the representations involved in propositional attitudes are processed in conformity with the sentential structure of their propositional content. Mental symbols behave like sentences. This ‘sententialist’ constraint on folk psychology, and the ‘sentential paradigm’ in cognitive psychology, are what eliminative materialism objects to (Churchland 1979, 125ff; Patricia Churchland 1986, 386ff).9

In his criticism of sententialism, Churchland points out that many important cognitive phenomena are not readily explained in terms of (para-)linguistic activity. Examples include the rotation of mental images, cognition in preverbal children, the intelligent behavior of animals and preverbal hominoids, as well as cognitive activity in the right, non-linguistic hemisphere in split-brain experiments. Other problems include the ‘linguistic catastrophe’ that, if sententialism is correct, must have occurred in the history of the species as well as that of individual language users. According to Churchland, these anomalies make it overwhelmingly clear that sententialism is false.

I believe the sententialist construction of intentionality, although largely correct as a diagnosis of much research in cognitive psychology (cf. Pylyshyn 1984; Fodor 1987 and 1990), has no direct basis in folk psychology itself.10 Let me take the case of preverbal humans and nonverbal animals as an example. Churchland’s argument appears to be as follows. Many creatures display intelligent behavior in spite of the fact that they cannot speak. But if they cannot speak to us, what reason is there to suppose that they can speak to themselves *in foro interno*, as folk psychology implies? It is thus impossible to explain the behavior of these creatures in terms of propositional attitudes. If we attribute propositional attitudes to them, as we sometimes do, this should be taken *metaphorically*. What we mean is that the child or animal, if it could speak to us like a normal human adult, would assent or dissent to our queries in the appropriate ways. Now, if children and animals have propositional attitudes
only metaphorically, then so have adults. For it is more probable that the information processing going on in young children and in animals is of the same kind as that in adult human beings; neither involves sentence-like representations. Therefore, sententialism is wrong.

This curious line of reasoning contains two objectionable steps. The first is an inference from lack of language to lack of intentionality. In everyday life, we attribute propositional attitudes to babies, dogs, cows, mice, frogs, and sometimes even to animals as unlikely as flies and spiders. Babies crave food and attention, dogs veer up in eager anticipation when you reach for their leash, cows obviously recognize the farmer coming down to milk them. The claim that these attributions are merely ‘metaphorical’ is absurd. There is no expectation that, if properly queried, your dog would answer correctly (no matter how much some people would like their pets to speak). Yet, in all these cases, the attributions are made for very good reasons. The behavior displayed by babies and beasts is such that it warrants the assumption that it is caused by internal representations of their environment, such as to account for their systematic discriminative reaction to distal factors, as required by definition (12).

The second step is a fallacious inference from sentence-described symbols to sentence-regimented symbols. From the fact that the content of the attitudes is described in terms of sentences, Churchland infers that they are linguistic in nature. This confusion of the form and content of symbols, though admittedly endemic among philosophers of mind (as we shall see in the following chapters), is not something to be held against folk psychology. A picture of victorious Bill Clinton has as its propositional content ‘that Bill Clinton has won the 1992 elections for the Presidency of the U.S’, yet it is not linguistic itself. Propositional attitudes in folk psychology may be relations between subjects and propositions, but they do not specify the form of the representation. In particular, they do not rule out the possibility that this form is neural, and that it can be scrutinized by neuroscience.

Once again, I want to plead for the utter philosophical innocuity of folk psychology. Earlier we have seen that folk psychology is in fact a relatively observational description of cognitive phenomena. Propositional attitudes specify the explananda rather than giving the explanations. From this it followed that the attitudes are also indispensable for any aspiring cognitive science. In terms of the above distinction, we can now state how these aspects are related: propositional attitudes specify the content of cognitive states (hence they are indispensable), while leaving their form indeterminate (hence they are
innocuous). In the next two sections, I turn to the methodological and empirical aspects of eliminative materialism to examine what happens when the propositional attitudes are systematically disregarded in cognitive science.

3. Methodological arguments

*Autonomous bottom-up approach defended*

In cognitive science, two broad research strategies can be distinguished, a bottom-up approach and a top-down approach. Top-down strategies start from some quite general account of the cognitive task to be explained, which is gradually decomposed into various routines and subroutines. The result is an understanding of the cognitive apparatus as a functional hierarchy of increasing complexity, until, eventually, it is broken down to its most elementary components (Dennett 1978; Lycan 1981 and 1990). Bottom-up strategies, by contrast, typically start from our knowledge of the elementary, functional parts of the nervous system, the individual neurons. They then try to establish how these are interconnected to form larger ensembles in more complex hierarchies, eventually working toward an explanation of cognitive achievements at a much higher level of analysis (Churchland 1988a, 96ff).

In practice, the prevailing method in cognitive psychology is almost exclusively top-down, whereas neurobiology typically avails itself of a bottom-up approach.¹¹ This methodological autonomy is philosophically vindicated by functionalism and eliminative materialism, respectively. According to functionalism, cognitive functions are indifferent with regard to their neural implementation. Hence, cognitive psychology should be conducted on a level of analysis that is free from any but the most marginal physical constraints. Eliminative materialism, by contrast, denies the validity of high level accounts of cognitive tasks as specified by folk psychology. Hence, the best methodology for cognitive (neuro)science is to disregard these accounts altogether, and start instead from the available scientific evidence on the hardware of cognition.

In this section I discuss three arguments in defense of an autonomous bottom-up approach. The first argument draws on the obvious fact that, in living things, cognition is somehow realized by the nervous system. Hence, knowledge of the structure of the nervous system is an indispensable asset for any cognitive science. In particular, cognitive science must be guided by ‘bottom-up’ knowledge of the brain’s functional atoms, the individual nerve cells or neurons. The nervous system is a structure of such intricate complexity that it
is hard to see *where else* to start but at the most elementary level. Cognitive neurophysiology, in this respect, resembles cybernetics in its early stages. Before a computer of any complexity could be built, it had first to be established what the functional properties of the available elementary circuits were. The next step was to demonstrate how these elements could be massively organized so as to perform basic calculations. Only then did it make sense to experiment with massively regimenting these calculations to form a virtual machine, capable of cognitive performance at a more abstract level. By analogy, neurobiology must begin with neurons, tracing their functional roles in hierarchies of increasing complexity, in order to eventually explain the more abstract, cognitive properties of the nervous system.

In this first argument two aspects should be distinguished. On the one hand, it offers a counterargument against functionalism’s defense of an exclusively top-down approach; on the other hand, it serves as a defense of an exclusively bottom-up approach, disavowing folk psychology as a suitable starting point for cognitive science. I have no quarrel with the first aspect of

![Figure 3.2: Levels of functional analysis](image)

The figure shows two basic approaches to cognitive phenomena, top-down and bottom-up, employing opposite methods of functional decomposition into increasingly simpler elements and functional composition into increasingly complex elements.
the argument, which, I believe, is a valid objection against functionalism. Considering the fact that biological intelligence is the prototype case of a cognitive system, which serves as a standard for determining what a system should be capable of in order to count as ‘cognitive’, it obviously follows that knowledge of the biology of cognition is indispensable. Of course, this is not to deny the possibility that nonbiological systems such as computers, or heterobiological systems such as Martians and other aliens, may be capable of cognitive activity of much the same kind as we are. It may well turn out that our notion of cognition should be loosened, and couched in more abstract, functional terms. That, however, is not something to be presupposed at the outset; rather, it is to be empirically discovered. This part of the argument, then, casts doubt on the methodological autonomy of top-down cognitive psychology.

As for the second aspect of the argument, I do not think it is convincing. Far from implying an exclusively bottom-up approach for cognitive science, it seems rather to invite bottom-up elements to be complemented with top-down elements, and vice versa, in a process of mutual conceptual and evidential interaction. Moreover, this suggestion is supported by the analogy with computer science. In modern computer science, the development of micro-architecture is continually constrained by top-down considerations (‘What abstract function do we want the machine to instantiate?’), while bottom-up constraints guide the development of new software (‘Which are the available architectures?’). Although software and hardware developers often work in separate departments, the rate of progress depends heavily upon their coordination and continuous exchange of information.

Functionalism

A second argument is directed specifically against the functionalist defense of the top-down approach. According to Churchland, functionalism is in reality only “a smokescreen for the preservation of error and confusion”, an immunization stratagem for saving bad theories from elimination (1981, 78ff). Churchland shows how the functionalist ploy might even have saved the ancient alchemistical four spirit doctrine. According to one school of alchemists, the observable properties of inanimate substances should be understood in terms of the four essences or spiritus residing in matter: the spirits of mercury, sulphur, yellow arsenic, and ammonia. Each of these spirits was held responsible for a characteristic syndrome of observable properties. The spirit of mercury, for example, which was present in metals, accounted for the power to reflect light, and to liquefy when heated. By the time the new corpuscular
chemistry of Lavoisier and Dalton became known, the four spirit doctrine had already been displaced. Yet, it might easily have been saved by a functionalist maneuver, so Churchland argues. Thus, having the spirit of mercury might have been reconstrued as a functional property of inanimate substances, defined as a certain disposition to reflect light, to liquefy when heated, and so on for the other properties the spirit used to account for. One consequence of this maneuver would be that the alleged science of alchemy is now able to define its object at a level that is distinct from and irreducible to that of corpuscular chemistry. The details of a substance’s corpuscular realization, the functionalist might claim, are largely irrelevant to our understanding of its functional properties, because many different corpuscular compositions can instantiate the same function. The lesson here is that functionalism could save any theory, even those that deserve to be eliminated as clearly as alchemy’s four spirit doctrine.

I believe that this objection is right in undermining the functionalist’s confidence in the well-known ‘multiple realization’ argument, which claims methodological autonomy for psychology by virtue of the ontological autonomy of functional states vis-à-vis their many possible realizations or instantiations (Putnam 1975, 408ff; Dennett 1978; Fodor 1974; for criticism, see Richardson 1979; Enç 1983; Churchland 1988a, 38-42). But this by itself does not mean that functionalism is completely wrong. One may reject the autonomy claim and still believe that mental states are, in an important sense, functional states. As intimated earlier, if we take as our starting point a study of the brain’s cognitive capacities, and try to graft them onto other systems, it may well turn out that our initial, relatively parochial and species-bound notion of cognition needs to be loosened in a functionalist sense (though this is an empirical question, not to be decided by a priori stipulation).

Does the alchemistical argument show that functionalism is wrong as a research strategy? The acceptability of functionalism in this sense obviously hinges on a prior diagnosis of folk psychology. If it is wrong to apply a research strategy such as functionalism to a bad theory such as alchemy, it does not necessarily follow that it is also wrong to apply it to a sound one. Any procedure that is fed the wrong input is bound to come up with the wrong results, but this is hardly the procedure’s fault. Although Churchland’s example seems to suggest otherwise, functionalist strategies are actually quite common in natural science, and, on the whole, perfectly innocuous. Thus, the relations between a gas’s functional properties at the phenomenological level (pressure, volume, temperature) can also be described at the corpuscular level,
in terms of statistical properties of the behavior of individual molecules. Superconductors have functional properties that can be studied at the phenomenological level as well as at that of their atomic structure. Notoriously, in biology we find a widespread use of various, interlacing, functional hierarchies, ranging from the levels of populations and organisms to those of cells and molecules. (The case of biology is particularly relevant here, as it is the discipline that is probably closest to cognitive science from Churchland’s point of view.) In sum, the issue is not whether the functionalist, top-down strategy is sound, but rather whether the phenomena it is applied to are robust. In classical cognitive psychology, these phenomena are specified in terms of the descriptive vocabulary of folk psychology. As for this starting point, however, we have already seen that there is, generally speaking, no reason for believing that it is completely mistaken.

Elimination and the wager

Perhaps the above conclusion about functionalism’s acceptability as a research strategy should be resisted, however. Although we have no reason to believe that folk psychology is completely mistaken, it is reasonable to assume that parts of it are. Suppose the ontology of folk psychology is partially wrong. A top-down analysis as described above would include not only the good parts but the bad parts as well, vesting them indiscriminately with an a priori assumption of scientific robustness. By the nature of its research method, functionalism is liable to carry these misconceptions further along, magnifying them at every step of the analysis. Thus, the small error at the beginning is turned into a big mistake toward the end. This is a risk we should protect ourselves from. One way to do so is to base our research on the methodological fiction that our common-sense view of the mental is completely wrong. As Churchland puts it,

“It is true that the bottom-up approach does not address directly the familiar mentalistic phenomena recognized in folk psychology, but that fact can be seen as a virtue of the approach. If the thumb-worn categories of folk psychology (belief, desire, consciousness, and so on) really do possess objective integrity, then the bottom-up approach will eventually lead us back to them. And if they do not, then the bottom-up approach, being so closely tied to the empirical brain, offers the best hope for constructing a new and more adequate set of concepts with which to understand our inner lives” (Churchland 1988a, 97).
Churchland’s line of reasoning here is strangely reminiscent of Blaise Pascal’s argument of the wager. In this famous argument, Pascal proved that it is always more profitable to believe that God exists than to believe that he does not exist. If, on Judgment Day, it turns out that God does exist, and the infidels are punished, one’s faith has not been in vain, and one is infinitely rewarded for having believed in God. If, on the other hand, it turns out that God does not exist, then the belief in his existence will have been a small effort, and one will have lost nothing. By analogy, in the case of cognitive science the question is whether the mind exists, as described by folk psychology. Churchland argues that one is always better off wagering that it does not exist.

Earlier in this chapter, I argued that folk theories are continuous with scientific theories in the same domain. Even as scientific theories mature and grow more independent, introducing new conceptual resources and identifying new phenomena and explananda, this continuity persists — otherwise these theories simply cease to be about the same domain as the prescientific ones. To the extent that neuroscience disregards the familiar folk psychological concepts, it fails to address the phenomena identified as cognitive. Although autonomous bottom-up research may discover many significant facts about action potentials, synaptic signal transmission and lateral geniculate bodies, it has nothing to say about perception, memory and mental representation, let alone about reasoning and motives. An approach that takes folk psychology serious is, in this sense, cognitive science’s best option faute de mieux. Notice, though, that this reply is aimed not so much at Churchland’s argument itself, as at the ideology behind it. In particular, I am not claiming that the argument is completely mistaken. As a matter of fact, I will endorse a weaker version of it myself in the next chapter, to the effect that the top-down approach must be applied carefully, not that it should not be applied at all.

Surprisingly, perhaps, this point is supported by the analogy between Churchland’s argument and Pascal’s proof of the existence of God. If God does not exist, but you have still wagered that he does, it will take a very long time before you can be certain that he does not exist. Doomsday simply will not come. Analogously, if propositional attitudes exist, but you have still wagered against them, it will take a very long time before the philosophical counterpart of Doomsday comes along. Only when the Peircean notion of an ideally matured neuroscience becomes reality, can we be certain that the attitudes exist. But again, this day will never come, for autonomous neuroscience cannot possibly mature. Paraphrasing Fodor (1980), no doubt it is all right to have a research strategy that says, ‘Wait a while’, but who wants to wait forever?
David Marr, the famous British mathematician and cognitive scientist, has pointed out that cognitive neuroscience must start from what he called a “computational theory” of the domain, which specifies the cognitive ‘problems’ that are ‘solved’ by the structure under investigation (Marr 1982, 19ff). Without such a prior notion of what is to be explained, neuroscience finds itself in a mere labyrinth of neural connections. As we have seen, the explananda of cognition are specified in the vocabulary of the propositional attitudes. In Marr’s terminology, folk psychology serves as the computational theory, or as the framework within which any computational theory must be set. This brings us to our next question: how successful has the autonomous bottom-up approach proved so far?

4. Empirical arguments

**Empirical decline of autonomous neuroscience**

It is difficult to give a general standard of empirical success, suited for objectively quantifying a theory’s degree of empirical progress or degeneration. Yet, in individual case studies it may often be possible to evaluate a theory’s empirical record in the light of its other achievements, most notably the strength or weakness of its conceptual and methodological resources. I think the case of eliminative materialism and the type of neuroscience it advocates illustrates this point. In the previous sections, I have argued that close consideration of eliminative materialism reveals a series of interlocking misconceptions and deficiencies. Should we now find that, in addition, the scientific program it advocates is virtually sterile, empirically speaking, then we have reason to believe something is wrong. A misguided method and a faulty philosophy conspire to make a lack of empirical success look extremely suspicious.

An example I take to be uncontroversial here is the classic work on the mammalian visual system, as conducted by Kuffler, Hubel and Wiesel in the 1950s and 1960s. In a well-known series of studies, these neurophysiologists investigated the way single cells in the mammalian visual system respond to simple light stimuli in their receptive field (see, for example, Bruce and Green 1985; Roth and Frisby 1986; Watt 1988). In this context, a nerve cell’s receptive field is defined as that part of the retina, stimulation of which causes changes in the electrochemical activity of the nerve cell.

In 1953, Stephen Kuffler had discovered that a certain type of cells in the retina, the so-called retinal ganglion cells, are selectively responsive to blob-like stimuli, formed by more or less circular boundaries between light and
dark (on-centre off-surround or off-centre on-surround). Subsequent discoveries were made by David Hubel and Thorsten Wiesel. Further up the visual pathways, in the primary visual cortex (area 17), they found cells that are most readily influenced by light/dark edges at particular orientations. Figure 3.3 shows a typical example of a cortical ‘simple cell’ of this kind, which responds most effectively to a vertically oriented boundary between light and dark.

Further research revealed the existence of various other types of simple cells in the visual cortex, as well as ‘complex’ and ‘hypercomplex’ cells. At each level investigated by Hubel and Wiesel, the receptive field properties appeared to elaborate on those at lower levels. Some cells were found to be sensitive to stimuli of particular shapes moving in particular directions, to bars, edges, or global features of wide receptive fields. For example, some cells resemble cortical simple cells in responding to a boundary at a particular orientation, but differ from them in that this boundary can be located anywhere

Figure 3.3: Electric activity of so-called cortical simple cells. The diagram shows the receptive field of a single cortical simple cell under various stimulus conditions (left), together with its recorded subsequent electrophysiological reactions (right).
within a very large receptive field, thus abstracting from the retinal position of the stimulus.

The results of this research were explained by the now classic feature detection model of visual information processing. According to this model, the mammalian visual system consists of a hierarchy of feature detectors, in which complex and hypercomplex cells process the information made available by cortical simple cells and retinal ganglion cells. The hypothesis is readily summarized in the following two claims.

13. Individual neurons in the visual system act as reliable detectors of properties of distal objects.
14. These neurons are connected hierarchically, such that detectors of relatively complex properties are fed with the output of detectors of less complex properties.

In spite of its promising start, research in the 1970s failed to come up with new insights in the postulated hierarchy of feature detectors. The program came to an empirical halt. In the words of David Marr,

“somewhere underneath, something was going wrong. The initial discoveries in the 1950s and 1960s were not being followed by equally dramatic discoveries in the 1970s. (…) The leaders of the 1960s had turned away from what they had been doing — Hubel and Wiesel concentrated on anatomy, Barlow turned to psychophysics, and the mainstream of neurophysiology concentrated on development and plasticity. (…) None of the new studies succeeded in elucidating the function of the visual cortex” (Marr 1982, 14-15).

In the light of our earlier results regarding the conceptual and methodological aspects of the unmitigated bottom-up approach, we can now give a rational reconstruction of why the program had to come to a halt. The problem to be solved can be summarized as follows:

15. How does the nervous system (and in particular the visual system) enable an organism S to form a reliable representation R of its environment, such that S can use R to efficiently organize its behavior?

The research method dictated by the bottom-up approach consists of essentially two instructions:
16. Find out which properties are detected by which neurons.
17. Determine the neural circuitry that enables the system to recognize increasingly complex properties.

Although these instructions seem fairly innocuous, they charge neuroscience with a task that is, in effect, impossible. Stochastically speaking, one is better off looking for a needle in a haystack.

A striking illustration of the complexity of the problem is the story of how Hubel and Wiesel made their famous discovery of the moving bar detector, as related by Bloom et al. (1978). Hubel and Wiesel had initially been looking for selective centre/surround sensitivity in the visual cortex, of the sort Kuffler had discovered in retinal ganglion cells, and they themselves had demonstrated in the lateral geniculate body (a neural connection between retina and visual cortex). In the cortex, however, no detectors of this kind were found (as we now know, because only layer IV of area 17, directly connected to the lateral geniculate bodies, contains cells of this sort). Yet, the microvoltmeter registered apparently spontaneous outbursts of activity in certain cortical cells, obviously unrelated to the centre/surround stimuli projected on the screen. The phenomenon defied explanation, until, eventually, it turned out that the cortex was responding to the vertical bar that moved across the screen when the projector slides were changed.

The lesson to be learned here should be clear. The search for the hierarchy of feature detectors has not made any progress because the problem to be solved, (15), is much too complex for the limited resources, (16) and (17), afforded by an exclusively bottom-up neuroscience. Waiting for chance discoveries is not a sound research strategy.

A further problem is illustrated in figure 3.4. The illustration shows a cortical simple cell that functions as a vertical bar-detector, (A), and one that functions as an edge detector, (B). In both cases, the cell’s physiological properties are such, however, that it cannot possibly distinguish between the different stimuli rendered in the picture; the net result of stimulation and inhibition, which determines the cell’s activity, is the same on either condition. As a consequence of this, the question which function is served by the cell does not have a straightforward answer if we consider only the input/output correlations between receptive field and individual cellular activity. From this viewpoint, the information carried by the cell’s activity is not that a vertical bar is present in receptive field R, but rather that the stimulus in field R is such that the net result of inhibition and stimulation has a certain value V. This may tell
us something about the intrinsic electrophysiological properties of the cell, but not about the cognitive function it subserves. Obviously, if this kind of indeterminacy occurs in cortical simple cells, one can readily imagine the situation to be even worse in complex and hypercomplex cells.

The example shows that a simple receptive-field-to-function inference, of the kind encouraged by eliminative materialism, is insufficient for explaining the cognitive relevance of neural activity. This demonstrates the ‘micromechanism’, so to speak, of an objection raised earlier in this chapter, namely, that eliminative materialism ‘changes the subject’ by failing to address the proper explananda of cognition.

Much the same conclusion can be drawn from a consideration of recent connectionist models of stereoptics, exemplified by the work of Lehky and

![Figure 3.4: Irresolvable ambiguity in cortical simple cells.](image)

The first diagram shows the receptive field of a single cortical simple cell (bar detector) under different stimulus conditions. A vertically oriented slit which is wider than the excitatory region of the cell’s receptive field (top left) evokes the same response as an oblique slit the same width as the excitatory region (top right). The second diagram shows the receptive field of a single cortical simple cell (edge detector). A low contrast edge at a vertical orientation (bottom left) and a high contrast edge at an oblique orientation (bottom right) evoke the same response. (After Bruce and Green 1985.)
Sejnowski on networks that compute shape from shading (Lehky and Sejnowski 1988 and 1990; Churchland and Sejnowski 1992, 183ff). Their aim was to build a simple feedforward network for computing the curvature of 3D bodies from 2D grayscale images. The neural net they used consisted of three layers of neuron-like units: an input layer of units with on-centre and off-centre receptive fields similar to cells found in the lateral geniculate body, an output layer of units jointly representing curvature, and an intermediate layer of ‘hidden’ units performing the actual computation. In the trained-up network, the hidden units appeared to have developed receptive field properties similar to those of simple cells in the visual cortex. Judging from their receptive fields alone, the hidden units seemed to be involved in edge-detection and bar-detection. Yet, their demonstrable and acquired function was to extract curvature from shaded images. As in the previous example, the conclusion should be that the cognitive function subserved by a unit cannot be determined by merely recording its receptive field properties. One needs to take into consideration the projective fields as well, that is, the way the unit’s output is processed at higher levels.

Although an exclusively bottom-up research strategy is ideally suited for simple feedforward models of feature detection, the limitations of the approach are rapidly beginning to show. When it comes to solving complex problems such as (15), the approach is clearly stretched beyond the reach of its means. In their evaluation of the experiment related above, Patricia Churchland and Terry Sejnowski conclude that

“a first-blush interpretation of single-cell recordings based solely on correlations between stimulus presented and response properties of the cells could in fact be highly misleading. Receptive field information alone is not enough to interpret a cell’s function. Given the favored assumptions behind ‘feature detector’ cells and exclusively bottom-up research strategies, this is a demonstration whose ripples extend a long way” (Churchland and Sejnowski 1992, 188).

It hardly needs pointing out that, if a problem-guided approach is indispensable for our understanding of the early stages of perception and other relatively peripheral processes, it is a fortiori indispensable for our understanding of central processes such as memory, learning, language, reasoning, or problem-solving.
5. Folk psychology explained

*Why folk psychology cannot be missed*

Against eliminative materialism, I have argued that our common-sense notion of the mental cannot be missed in cognitive science. It may in due course come to be *revised*, but it cannot be dismissed wholesale. The reason for this, or so I argued, is that folk psychology provides the relatively observational vocabulary in terms of which cognitive phenomena are specified; to disregard it is to change the subject. In other words, I subscribe to the “abstract” and “faintly stipulative” character of folk psychology, as Churchland himself has called it (1981, 77ff). We have seen various ramifications of this view. Propositional attitudes specify the *explananda* rather than giving the *explanations*. Methodologically speaking, folk psychology serves as the framework in which, as Marr used to call it, the *computational theory* of the domain is couched. Without such ‘top-down’ guidance, cognitive (neuro)science is faced with a problem it cannot possibly solve, inevitably leading to empirical and conceptual stagnation. As a way out of this quandary, I urged the need for continual *conceptual interaction* between folk psychology’s descriptive resources and the developing explanatory apparatus of cognitive science, including neuroscience and connectionism.

An important topic in the above discussion was the notion of intentionality as purportedly endorsed by folk psychology. Thus, Churchland submitted that the incommensurability between folk psychology’s intentional categories and the framework of physicalist science prevents the latter from ‘benefiting’ by any conceptual interaction with the language of propositional attitudes. Against this claim, I argued that it relies on a notion of intentionality that is considerably richer than folk psychology is committed to. Following up on earlier suggestions regarding the theoretical innocuity of folk psychology, I proposed an alternative, ‘minimalist’ reading of the propositional attitudes and their intentionality, which is committed neither to abstruse mental metaphysics, nor to a sententialist construction of mental symbols.13 In this final section, I want to expand somewhat on this latter issue, in preparation of our discussion of mental symbols in subsequent chapters. I will do so by sketching in rough outline a coherent view of the relation between folk psychology and cognition, which weaves together various strands of argument encountered in this chapter. In particular, I will explain the sense in which the propositional attitudes *define* the explananda of cognition.
A dual aspect theory of the propositional attitudes

In my discussion of Churchland’s objection to sententialism, I argued that the propositional attitudes specify the content of cognitive states (hence they are indispensable), while leaving their form largely indeterminate (hence they are theoretically innocuous). I now want to enlarge upon this distinction between form and content. I think that, with regard to the attitudes, two aspects of quite different philosophical pedigree should be sharply distinguished. On the one hand, assuming we are ‘intentional realists’ (cf. Fodor 1985), the states identified in propositional attitude attributions are supposed to be real states of real subjects. The question facing any prospective cognitive science in this respect is whether the attitudes can be naturalized, that is, whether their reality can be vindicated, and their nature be explained, by cognitive science. Accordingly, I refer to this aspect of the attitudes as the real or natural aspect.

18. Natural aspect: propositional attitudes specify real states and processes of cognitive subjects.

In addition to this natural aspect, propositional attitudes have a second aspect that is easily overlooked from a naturalist point of view. I call this second aspect logical or epistemic.

Figure 3.5: Two aspects of the propositional attitudes

Attribution of propositional attitudes identifies the epistemic content of cognitive processes rather than their specific natural form (left). Cognitive processes involve two distinct aspects, logical and natural (right). In the approach sketched in this chapter, the cognitive subject serves as the link between these two aspects: its knowledge, identified in terms of propositional attitude attributions, is both logical and real.

Notice that, in formulating (18), I assumed a realist perspective merely for convenience of exposition: whether or not you accept (18), you will have to give some account of why propositional attitudes do or do not exist (and if you think they do not exist, you owe an additional explanation of folk psychology’s instrumental success at explaining and predicting behavior); hence, a naturalistic account.

Even if you are an instrumentalist or eliminativist with regard to propositional attitudes, and hence deny proposition (18), you are still bound to the epistemic aspect captured by proposition (19). It may be possible to deny that the attitudes exist, but it is difficult to see how anyone could deny that they describe our relation to the objects of cognition. Let me give a simple example here. Recall the belief attributed to Luke Skywalker, namely, that Darth Vader is his father. In its natural aspect, this belief is put forward as a candidate entity or state, such that there is something inside Luke’s head that can be properly identified as this entity or state, and that is causally responsible for his overt behavior. This may or may not be true. Now consider the logical aspect of Luke’s belief. It specifies a relation between Luke Skywalker and various entities and situations in his environment, both past, present, and future. This is the sense in which his belief is about the black-robed entity referred to as Darth Vader, and about fathers in general, and about his own father in particular. There is nothing ontologically suspect about these entities and situations, nor about Luke’s relation to them. They are not mental, nor even are they inside the head; they are simply objects in Luke’s environment, or interactions between such objects. The important point is that, in attributing to Luke the belief in question, part of what we do is to specify a logical or epistemic relation between Luke and these objects. This relation is cognition. To deny this aspect of cognition is to deny cognition itself; which is absurd.

In their natural aspect, the attitudes are supposed to reveal something of the real internal structure of subjects. They are supposed to specify the nature of the subject’s relation toward internally stored symbols, as well as the nature of those symbols themselves. Churchland’s claim that propositional attitudes are committed to a view of the symbols as internally stored sentences obviously falls within the province of this first aspect. Against this claim, I submitted that the notion of mental symbols as used in folk psychology is much less sophisticated. In their natural aspect, propositional attitudes refer to the form of the
symbols employed—a form folk psychology leaves largely indeterminate. Yet, even if we were to accept Churchland’s diagnosis of the first aspect of the attitudes, it would still leave the second aspect completely intact. This second aspect specifies the contents of cognitive states, the information the cognitive system is supposed to be processing. Crudely put, and somewhat paradoxically, folk psychology offers not so much a theory of mind, but primarily a theory of what mind is about. In this sense, to deny folk psychology is not only to deny the mind, but also to deny the world; which is absurd.

The ideal name for the outlook presented here is obviously the dual aspect theory of the propositional attitudes. Unfortunately, however, this name, or one very similar to it, has already been taken by a well-known and quite different doctrine in the same field. I am referring to the ‘double aspect’ theory of mind proposed by Sir Peter Strawson and others (Strawson 1959, ch. 3; Meijsing 1986, ch. 6), a doctrine that is metaphysical in character as opposed to my epistemic concern. For ease of reference, I will nonetheless accept the name as an epithet for my position. Within the confines of the present study, confusion is easily avoided. Henceforth, if I use the expression ‘dual aspect theory’, I refer to the epistemic theory outlined above, unless stated otherwise.

The distinction invoked by the dual aspect theory should not be confused with a more familiar distinction that runs partly parallel to it, namely, that between the attitudinal and the propositional aspects of propositional attitudes. On this distinction, folk psychology’s repertoire of epistemic attitudes, such as ‘believing’, ‘desiring’, ‘remembering’ and ‘perceiving’, describes the prescientific functional organization of the mind. In folk psychology, the corresponding functional compartments of the mind are specified essentially in terms of their i/o-control and mutual interrelations. The perception department, for example, is specified in terms of a set of characteristic relations between observable circumstances and dispositions for belief fixation, while the ‘belief box’ is identified, inter alia, as a readiness to respond that X when queried whether X. The propositional aspect of the attitudes, on the other hand, serves a quite different purpose. Seen from the perspective just sketched, it specifies not the functional organization of the mind, but rather the information-bearing items that are supposed to be processed by it. This, I take it, is roughly the view of folk psychology endorsed by Jerry Fodor and many other philosophers (cf. Fodor 1987; Pylyshyn 1984).

The second distinction is different but in part parallel to the first. It is parallel to the extent that it draws attention to the propositional content of cognitive states as being importantly different from the functional organization of
the mind. But here the comparison ends, for the logical aspect of the first distinction also extends beyond the propositional aspect of the second. Similarly, the natural aspect covers not only the attitudinal organization of the mind, but also the form of the symbols it uses. Thus, whether or not propositional attitudes commit us to the existence of separate ‘belief boxes’ and ‘memory banks’ (a natural question), they do commit us to cognitive relations of various sorts (a logical question). Similarly, whether or not the attitudes commit us to sentence-like internal symbols, they do commit us to specify the content of cognitive states in terms of the descriptive vocabulary of the propositional attitudes.

Although the second and more popular distinction may not really be false as far as it goes, I believe it tends to suggest a picture of cognition that is badly misleading. It invites us to think of cognition as essentially an autonomous process going on entirely inside the head. Concepts, thoughts and other inhabitants of the mental are readily seen as self-identifying items that pop in and out of belief boxes, and that are fed from one function to another. From this purview, as Fodor has urged, the items are quite naturally analyzed as endowed with a proper combinatorial, syntactic structure, which determines their intrinsic, semantic content. This notion of intrinsic content we shall meet again in the context of various discussions in subsequent chapters.

What is next?

In the following two chapters, I turn to a discussion of connectionist theories of mind. As intimated earlier, connectionism is put forward by Churchland and others as a radical alternative to cognitive psychology, while it is claimed by critics to be unable to explain cognition. Critics of connectionism typically endorse both Churchland’s incommensurability claim and the objection of explanatory irrelevance raised against eliminative materialism in this chapter. I will address the question whether the charges leveled against connectionism are justified.
Chapter five

Qualia and epistemic content

“How sense-luscious the world is”, Diane Ackerman marvels in her *Natural history of the senses* (1990). Sensations of all sorts, in their countless hues and combinations, color, taste, smell, touch, sound, and other feelings: they delight and define our sense of world and consciousness alike.

“There is no way in which to understand the world without first detecting it through the radar-net of our senses. (...) Our senses define the edge of consciousness, and because we are born explorers and questors after the unknown, we spend a lot of our lives pacing that windswept perimeter: We take drugs; we go to circuses; we tramp through jungles; we listen to loud music; we purchase exotic fragrances; we pay hugely for culinary novelties, and are even willing to risk our lives to sample a new taste” (Ackerman 1990, xv).

In the philosophy of mind, the qualitative contents of sensations are technically known as *qualia*: the ways things appear to us in phenomenal consciousness—from the fragrance of coffee in the morning, to the soft, cool pressure of fresh sheets at night.

Qualia are dangerous things, and not only because we are prepared to risk our lives to taste a new sensation. Also from a philosophical point of view, qualia exude a precarious charm. Thus, their epistemological format bears the immediacy and indubitability sought by Descartes: to have them is to know what they are. Moreover, their ontological stature resembles what the Bishop of Cloyne claimed for all being: their existence consists in their being perceived by us. Finally, qualia possess the ineffable subjectivity captured by the Scholastic adage, ‘de gustibus non est disputandum’: they are strictly private, and lie beyond the reach of all conceptual comparison. All of these characteristics mark qualia as dangerous things, not to be approached without a certain apprehension.

If we are looking for something to eliminate from the realm of science or folk psychology, qualia seem a particularly good place to begin. Their condi-
tions of identity are extremely vague, as has been repeatedly demonstrated by Daniel Dennett and others in carefully chosen thought experiments (Dennett 1990; 1991). Considering the fact that there can be ‘no entity without identity’, as Quine impressed on us, qualia had better be eliminated from the ontology of mind. At the same time, however, qualia seem especially bad candidates for elimination. Their phenomenological reality is beyond dispute; nothing is more clearly and more vividly present to the conscious mind than qualia. By this token, it is often argued that the existence of qualia demonstrates the inadequacy of physicalist accounts of the mental. The subjective, qualitative nature of qualia, intuited ‘from within’, is believed to withstand all third-person explanations of man and his place in the world.¹ A dilemma presents itself. Physicalists will be inclined to eliminate qualia because they do not fit in with our scientific worldview, while opponents will be inclined to denounce that very same worldview because it is unable to account for qualia. To break the deadlock, we must show how qualia can be stripped of the philosophical mystery surrounding them, while at the same time adding to their sense of naturalness and utility.

I believe that qualia can indeed be meaningfully incorporated in a physicalist theory of mind. In order to show how this can be done, I take my lead from Churchland’s treatment of qualia, and discuss a number of important objections raised against it. Even though Churchland’s position may be mistaken in important respects, I believe that, with the proper corrections and additions in place, it can be fruitfully used for advancing our understanding of neural epistemics in the field of sensations.

1. Qualia disqualified?

Two fundamental ambiguities

The claim that perception is a causal process has been made by a number of philosophers in the recent past (see, for example, Churchland 1979; Dretske 1981; Fodor 1987). Causal theories of perception are based on the idea that the nervous system functions as essentially a measuring instrument, which is used by the organism for the reliable detection of properties of distal objects. Objects cause detection states, which can then be used by the organism to infer properties of the distal object. As Churchland puts it,

“Perception consists in the conceptual exploitation of the natural information contained in our sensations or sensory states” (1979, 7, italics mine).
Causal theories raise many questions, both at the object end and at the subject end of perception, on which I shall have more to say in subsequent chapters. For the moment, my plan is to concentrate on the subjective target where sensations and their qualitative contents are located. As indicated above, sensations are marked by an introspectable, phenomenal content, or \textit{quale} for short. Now, Churchland’s proposal with regard to qualia is simply to \textit{identify} them with states of the detection device—hence the disjunctive clause in the quotation, “sensations or sensory states”.

The alluring clarity of this proposal is spoiled by two fundamental difficulties. First, a deep philosophical controversy lurks beneath its flawless surface. The identity of \textit{sensations} is primarily of an epistemic or semantic nature. Sensations are picked out by their specific qualitative content. \textit{Sensory states}, by contrast, enjoy a causal identity. More in particular, their identity is established in neurophysiology: being states of the sensory apparatus, they are picked out in terms of their role in the physiology of perception. In other words, the taxonomy of sensations is \textit{semantic}, whereas that of sensory states is non-semantic, or, as it is also called, \textit{syntactic} or \textit{formal} (a terminology that will be discussed in more detail in chapter seven). Can these taxonomies be equated, as Churchland suggests, or is the equation simply confused?

The second problem is an ambiguity in Churchland’s attitude toward mental phenomena in general, and toward qualia in particular. Strictly speaking, qualia are products of folk psychology, to which Churchland’s general attitude tends to be eliminative. As we saw in chapter three, eliminative materialism is not a sound position. Now, in the case of qualia, Churchland himself seems to hesitate, too. Sometimes his position is still distinctly eliminative, but at other times he appears to favor a \textit{reduction} of qualia. The eliminativist view tends to prevail in particular in Churchland’s earlier work (1979; cf. 1981), where much emphasis is on the fact that the qualitative identity of sensations is \textit{redundant}.

“The intrinsic qualitative identity of one’s sensations is irrelevant to what properties one can or does perceive the world as displaying. (…) Sensations are just \textit{causal} middle-men in the process of perception, and one kind will serve as well as another so long as it enjoys the right causal connections. (…) In principle they might even be \textit{dispensed} with, so far as the business of learning and theorizing about the world is concerned. As long as there remain systematic causal connections between kinds of states of affairs and kinds of singular judgements, the evaluation of theories can continue to take place” (Churchland 1979, 15).
Passages such as these readily suggest a selective application of elimination and reduction: sensations are eliminated, while sensory states can be retained as candidates for reductive explanation in terms of neuroscience. Yet, this does not seem to be Churchland’s intention, as is clear from the frequent appeal, in particular in his more recent work, to the possibility of reducing the qualitative aspect of sensations. Characteristic of this tendency are passages like the following.

“The objective qualia (redness, warmth, etc.) should never have been ‘kicked inwards to the minds of observers’ in the first place. They should be confronted squarely, and they should be reduced where they stand: outside the human observer. (...) If objective phenomenal properties are so treated, then subjective qualia can be confronted with parallel fortrightness, and can be reduced where they stand: inside the human observer” (Churchland 1985a (1989a, 57)).

“The ‘ineffable’ pink of one’s current visual sensation may be richly and precisely expressible as a 95Hz/80Hz/80Hz ‘chord’ in the relevant triune cortical system. The ‘unconveyable’ taste sensation produced by the fabled Australian health tonic Vegamite might be quite poignantly conveyed as a 85/80/90/15 ‘chord’ in one’s four-channeled gustatory system (a dark corner of taste-space that is best avoided). And the ‘indescribable’ olfactory sensation produced by a newly opened rose might be quite accurately described as a 95/35/10/80/60/55 ‘chord’ in some six-dimensional system within one’s olfactory bulb” (Paul Churchland 1986 (1989a, 106)).

One of my aims in this chapter is to track down the causes of these ambiguities. I think the wavering attitude toward elimination or reduction of qualia puts the finger on a difficulty with regard to mental content. An important clue to the solution of this problem can be found in a related idea of Churchland’s, namely, the claim that perception is plastic: that our very perception of the world may change as a result of a change of theory. I will first explain this claim itself, and then trace its ramifications with regard to qualitative content. Notice, though, that it is not my aim here to define or defend Churchlandian orthodoxy. As a matter of fact, I will suggest some important corrections and additions to what is probably Churchland’s own position. My purpose is rather to advance the scope of a general theory of neural epistemics, of the kind introduced in the previous chapter.
Empirical realism and theoryladen observation

Before examining the doctrine of plasticity of perception in more detail, I need to say something about the philosophical background from which it has emerged. Three basic tenets should be mentioned here, which are endorsed, in one form or another, by many contemporary philosophers of science: empiricism, scientific realism, and a network theory of meaning. They represent basic convictions in the fields of epistemology, metaphysics, and philosophical semantics, respectively. The question is whether this theoretical troika is consistent.

According to Churchland’s version of the network theory of meaning, the semantics of all sentences is determined in part by a sentence’s role in the overall system of beliefs in which it occurs. This claim is, in effect, the semantic counterpart of the evidential holism introduced in chapter three. There we saw that evidential holism applies to observation sentences as well as to theoretical sentences. The same is true of semantic holism. Not only is the meaning of theoretical sentences dependent upon the global set of theories in which they figure, as well as on observation reports, but also the meaning of observation sentences themselves is radically theoryladen. Some versions of holism allow for a measure of semantic and evidential autonomy for observation sentences, thus restricting the scope of theory-dependency. A typical example of this is the position of Quine (1970 and 1990).

A second recurring theme in contemporary philosophy of science is empiricism, understood as the claim that observational evidence is of prime importance to our knowledge of reality. Our senses convey to us information about the observable properties of the world. It is this sensory information that drives our understanding of reality, from the simplest observation reports to high-leveled explanations in quantum physics. Observation remains a constant constraint on all theoretical description and explanation.

The third principle to be mentioned here is that of scientific realism, the claim that all entities, properties, processes and events referred to in scientific explanations are vested with a prima facie presumption of existence. This claim applies to observable entities as well as to so-called ‘unobservable’ or ‘theoretical’ ones. Together, the second and third principle amount to a position that may properly be called empirical realism. One of the consequences of empirical realism, as I shall use the term here, is that observation sentences reliably and immediately report on real properties of distal objects. The terms used in observation reports, as well as those used in theories, refer directly to the real world; or, in more classical terms, it is reality itself that is
the primary object of knowledge. Thus understood, empirical realism sharply contrasts with a view on which mental representations, such as sensory impressions or sensations, are the primary objects of knowledge. According to the realist alternative, no intermediate mental content functioning as the primum notum is needed.

Each of the above principles is equably desirable from a philosophical point of view. Yet, the attempt to combine them poses severe difficulties (Philipse 1990; cf. Sleutels and Corbey 1992). If, as required by semantic holism, the meaning of observation sentences changes as a function of the observer’s background theory, then obviously this meaning is not determined directly by the object itself, but rather by the internalized theory about that object. What we observe is dictated more by our theory than by the world—a consequence which threatens to undermine scientific realism. Moreover, semantic holism entails that there is no principled distinction between observation sentences and theoretical sentences. But this means that the distinction between ‘observable’ and ‘theoretical’ entities starts to collapse: arguably, all entities are equally ‘unobservable’—a consequence that cuts at the root of empiricism. One way to avoid this difficulty, which has traditionally received much attention, is to appeal to the mental processes purportedly involved in observation and theory. Stated in rough outline, the idea is that one kind of mental representations (namely, ‘percepts’), which are characteristic of observation, occur at the earliest stages of cognitive processing, whereas another kind of mental representations (namely, ‘concepts’), which are characteristic of theoretical understanding, enter only at a later stage. This solution will not do, however. Although it successfully vindicates empiricism, it blatantly contradicts the key claims of holism as well as those of direct realism.

**Plasticity of perception**

A particularly radical version of the theoretical trojka is endorsed by Paul Churchland (1979; 1985b; 1988a; 1988b; cf. Philipse 1990, 129-146). What sets his position apart from that of many others is the fact that it explicitly explores the possibility of aligning the three principles in their radical form. To this end, an additional claim is introduced, namely, the principle of plasticity of perception, which is based on a descendant of the representational empiricism described in the previous paragraph. Churchland’s theory of perception draws on a distinction which reminds at first sight of that between percepts and concepts. Thus, at the outset of this section, we saw a distinction be-
between, on the one hand, the ‘sensations or sensory states’ that carry information about the environment, and, on the other hand, the representations that ‘conceptually exploit’ this information. Hence, perception consists in sensations being regimented by concepts, and eventually leading to the production of overt behavior, including the uttering of observation sentences. To this more or less traditional view is added an all-important new claim, however: that observation is malleable by theory. New theories may lead to new perceptions.

Stated in this form, the principle of plasticity is badly clouded by ambiguity. What exactly is the subject of this alleged plasticity? Two possibilities suggest themselves here: observation sentences and sensations. Obviously, these two interpretations yield entirely different varieties of perceptual plasticity: ‘linguistic’ or ‘judgmental plasticity’ on the one hand, and ‘sensational plasticity’ on the other hand. Observation sentences are spontaneous, non-inferential statements about distal factors; even if they change, sensations may stay the same. As a matter of fact, this is generally understood to be what happens when you switch to a foreign language. Seen from this angle, the possibility of judgmental plasticity seems to be real but rather trite (although I shall shortly argue that it is far from trivial from a philosophical point of view). Far more spectacular is the possibility of sensational plasticity. Sensations are the subjective counterparts of our spontaneous observation reports, identified in phenomenal consciousness by their qualitative contents. If they change due to a change of theory, then so will our qualia; the world will, quite literally, no longer look the same to us.

Churchland himself explicitly distinguishes between the two varieties of perceptual plasticity. As he puts it in his commentary on Fodor (1988):

“My own 1979 position (…) simply assumes the generally constant character of our sensory responses to the environment. The plasticity that excited me there was confined to the conceptual frameworks within which we make our judgmental responses to the passing contents of our sensory manifold. (…) To be sure, sensational plasticity would constitute and additional argument for the plasticity of perception. (…) And I (…) am now willing to defend it vigorously” (Churchland 1988b (1989a, 277-278)).

Plasticity of observation reports has come to be known as weak plasticity, while plasticity of sensations is now better known as strong plasticity. I will follow usage in this respect. Merely to have different names for different
doctrines, however, is no sufficient guarantee that they will not be conflated all the same. This question is particularly important if we want to know whether or not a species of the naturalistic fallacy is committed here, as explained in the opening chapter. Thus, critics have argued that Churchland’s attempt to steer clear from the naturalistic fallacy proceeds by systematically ambiguating between eliminativism and reductionism (Fodor and Lepore 1992), or between weak and strong plasticity (Fodor 1988), or both (Philipse 1990). I presently turn to an examination of these charges, with particular attention to the status of qualia in Churchland’s theory. First, however, I want to introduce some new terminology to facilitate my exposition in this chapter.

Qualia: weak, strong, superweak, or superstrong?

My first point of reference is the classical empiricist theory of perception, of the kind proposed by John Locke (1670), and criticized by Kant (1781) and Sellars (1963, 127-196; see also chapter one, above). According to Locke’s version of representational empiricism, what is immediately given in perception is not the distal object itself, but rather the qualitative, sensory impression it makes on consciousness. The subject is primarily conscious of a mental object, namely, of his own internal state, for example of an impression of red. Mediated by this impression, he may then come to (re-)construct a theory of the impression’s distal causes in the extramental object of perception. Locke’s theory of perception exemplifies one possible way to understand qualia, namely, as endowed with an epistemic identity of their own, which is simply ‘read off’ by the conscious subject. For ease of reference, I call this type of qualitative content superstrong.

1. Superstrong qualia =\text{def} \text{qualitative states of phenomenal consciousness that determine the meaning of observation sentences.}

Being epistemically active ‘givens’ in the sense criticized by Sellars, superstrong qualia readily invite the naturalistic fallacy. Hence, there is every reason to reject them. If they are disavowed, however, a new problem presents itself, for now it is no longer clear in what sense sensations can still be said to be relevant for perception at all. If the quale does not determine what is perceived, how can it determine that we perceive? Apparently, it ceases to be part of the essence of perception. Rejection of superstrong qualia thus seems to lead to a view on which qualia become superweak, in the sense recorded by the following definition.
2. Superweak qualia $=_{\text{def}}$ qualitative states of phenomenal consciousness that are not irrelevant for perception.

Churchland explicitly subscribes to Sellars’s criticism of the naturalistic fallacy, as is clear from his remark that “sensations themselves are not yet truth-valuable or semantically-contentful states”. They belong to “the wrong logical space: it is only an observation judgment, or belief, or report that can be logically consistent or inconsistent with any theory” (Churchland 1988b (1989a, 267-268 and 277)). Yet, it does not necessarily follow that sensations must be superweak if they cannot be superstrong. As intimated earlier, Churchland’s position with regard to qualia tends to be ambiguous: sometimes he stresses the fact that they are redundant (“they might even be dispensed with”), while at other times pressing the idea that they can be reduced. Also, we saw just now that Churchland uses two different varieties of perceptual plasticity, which each in their own way directly affect the status of qualia. To describe Churchland’s idea of qualia, I introduce two new notions here: weak qualia, which are ruled by the principle of weak plasticity, and strong qualia, which are ruled by the principle of strong plasticity. Stated more circumstantially, we may draw up the following definitions:

3. Weak qualia $=_{\text{def}}$ qualitative states of phenomenal consciousness whose qualitative identity is independent of their conceptual exploitation.
4. Strong qualia $=_{\text{def}}$ qualitative states of phenomenal consciousness whose qualitative identity is a function of their conceptual exploitation.

In terms of these four definitions, the objections against Churchland’s plasticity thesis can now be succinctly stated as follows. Critics argue that the distinction between weak qualia and strong qualia is doomed to collapse into either superweakness or superstrength; hence, all qualia are either cognitively irrelevant (superweak), or they commit the naturalistic fallacy (superstrong). The following sections take a closer look at this dilemma, starting with the case of weak qualia. I will argue for a reductive interpretation of weak qualia as well as strong, whilst at the same time trying to explain the sense in which qualia, both strong and weak, are ‘redundant’ from an epistemological point of view.
2. Weak qualia examined

The main argument advanced in behalf of weak qualia is that of different modalities leading to essentially the same sensation. After explaining the argument in some detail, I will turn to the criticism it has drawn from Herman Philipse in his recent paper on the philosophy of Churchland (Philipse 1990). Although I think Philipse’s objections are invalid in important respects, they may show the way to a more adequate understanding of the notion of weak plasticity. In particular, I want to investigate the sense in which superstrong qualia are claimed to account for the element of constitution of perception, in a way that weak qualia are purportedly unable to emulate.

The argument from different modalities

Churchland’s principle of weak plasticity entails that even observation terms such as ‘red’, ‘cold’ and ‘hot’ refer directly to the objective properties of the distal object that caused the corresponding sensations in the subject, rather than referring to these sensations themselves. In support of this claim, Churchland argues that different causal routes may lead to relevantly the same perception (1979, 7-14). For example, in human beings, heat and cold are perceived by means of a tactile or bodily sense for temperature; other species, such as the rattle snake, have a separate sense for electromagnetic radiation in the infrared part of the spectrum (see, for example, Newman and Hartline 1982). Now, imagine a society of hominoid beings that are like us in all respects, except for the fact that their retinas are sensitive to radiation in the far infrared—plus whatever additional modifications to eyeballs and/or lenses might be needed to allow them to visually detect heat. Considering that these hominoids resemble us so closely, it is plausible to suppose that they see a hot object in the same way as we see an incandescent white object. Moreover, like ourselves, these creatures use language to exchange observation reports. Now, if asked how to understand a hominoid’s report of seeing a very hot object, we are not even tempted to translate it as, ‘He sees a very white object’, although his sensation is, by hypothesis, like our sensations of white. Rather, we immediately assent to an object-guided translation of his report as, ‘He sees a very hot object’.

“On viewing a very hot object they have what we would describe as a sensation of an incandescent white object, and on viewing a very cold object they have what we would describe as a sensation of a black object,
and so on. They, of course, describe these sensations quite differently—as sensations of heat, of coldness, and so on” (Churchland 1979, 9).

Summarizing, we find the following claims to be intuitively plausible. Speaking objectively, in terms of distal causes, the heat that is seen and the heat that is felt are the same. Speaking subjectively, in terms of qualitative sensations, they are different. Speaking intersubjectively, in terms of linguistic behavior, they are the same again by virtue of the favored, object-guided translation.

The argument from different modalities successfully establishes the principle of weak plasticity. It succeeds in bridging the gap between network semantics and empirical realism, as is clear from the following consideration. The meaning of observation sentences (for example, ‘This is hot’, as uttered by infrared observers), is determined directly by the observed distal properties, as required by empirical realism. At the same time, however, it is also a function of the observer’s theory, as required by network semantics. In the example given here, the relevant theory of temperature is shared by hominoids and humans, which explains why their observation reports mean the same thing as ours. The important point here is the fact that weak qualia themselves are largely irrelevant in the process of perception: they determine neither what is perceived, nor how it is described.

A fundamental objection

In a critical analysis of Churchland’s philosophy by Herman Philipse (1990), the above conclusion has been challenged. Philipse draws attention to the fact that the argument from different modalities assumes that different modalities lead to qualitatively different sensations. Now, “if the way the world appears to the hominoids were not determined by their sensations”, so Philipse argues, “it would not make sense to consider the possibility of a sensation-guided translation, albeit as a possibility which is eliminated (...). In other words, it is essential to Churchland’s argument that his biological fiction presupposes the very variety of the representative theory of perception which is rejected once the argument has done its service” (Philipse 1990, 141, emphasis added). As is apparent from the context, the variety of the representative theory of perception allegedly presupposed by Churchland is Lockean empiricism.

At first sight, this objection seems just extravagant. If it is shown that qualitative contents are irrelevant to perception, it surely does not follow that
they are essentially presupposed by the theory. Notice that, ontologically speaking, Churchland’s view of weak qualia has a tendency toward epiphenomenalism, or is at least compatible with it. The subjectively introspectible quale is perfectly inert: even without it, all systematic connections between environment, perceptual processing, and behavior would remain intact. The causal inertia of weak qualia implies that they can do no epistemic service whatsoever: they do not determine what we perceive, nor what we will say about it. In this sense, Churchland can claim that they “might even be dispensed with, so far as the business of learning and theorizing about the world is concerned” (1979, 15).

Yet, there is more to the objection than meets the eye. In my explanation of weak plasticity, the causal inertia of weak qualia has become so complete, that they are in danger of becoming superweak. This is probably the true worry behind Philipse’s enthymeme. If the argument is elaborated in this sense, it takes the form of a dilemma: weak qualia must be either superstrong or superweak. If they are superweak, they are no longer constitutive of perception, as indeed seems to be the case. Hence, Churchland misses out on an important aspect of perception, as opposed to the physiology of neural processes. On the other hand, should weak qualia turn out to be superstrong, then the principle of plasticity goes by the board: by definition, superstrong contents of sensation determine the meaning of observation terms. In sum, the notion of ‘weak qualia’ seems to be an impossible hybrid, which wants to have the best of both worlds: from superweak qualia it takes the causal inertia, and from superstrong qualia their relevance to perception.

In this modified form, there is much to be said for the objection against weak plasticity. It is probably correct as a diagnosis of why Churchland uses the wavering disjunctive, ‘sensations or sensory states’, in his definition of perceptual processes (Philipse 1990, 157-158). Also, it may be one of the reasons why Churchland has apparently come back from his initial eliminativism and, in his more recent work, has begun to explore the possibilities of reducing sensations to neural states. Finally, considerations such as these may have motivated the introduction of strong plasticity of perception. Yet, I do not feel comfortable with the premises on which the objection is based. In particular, it is very difficult to explain the sense in which qualia are required to be constitutive of perception. The objection rests on the assumption that only superstrong qualia can meet this requirement. I think this assumption is too fastidious, and that it leads to absurd consequences. Let me explain this by trying to apply it to Churchland’s original example. The infrared-
sensitive hominoids, so the (super-)strong version of the argument will read, have basically the same sensations as we do, namely, sensations determined by their intrinsic contents of ‘white’. Unlike us, however, the hominoids interpret these contents as representations of hot objects. The immediate perception of ‘white’-qualia leads them to infer heat, and leads us to infer whiteness. The qualia are the same in both cases, but the perceptual judgments are different. Apparently, the element of constitution of perception is contained in the introspective consciousness of ‘pure’, qualitative ‘givens’ prior to the judgment.

Breaking the dilemma of superstrong and superweak qualia

The consequences of the above view seem highly undesirable to me. Let me mention three fundamental difficulties here. In the first place, the theory is circular as an explanation of perception, for it evidently presupposes the existence of a homunculus capable of perceiving, interpreting, and acting on qualia. Moreover, the content of qualia that is purportedly ‘given’ to consciousness is either arbitrary (for why should it be called ‘white’ rather than ‘hot’?), or indefinitely disjunctive (‘hot or white or … ’). This excavation of superstrong content prevents it from doing any epistemic work at all: how could such indefinite content ever serve to identify the ‘real’ perception? Finally, the idea of superstrong content is contradicted by the phenomenology of perception: no introspective consciousness of pure contents precedes my exterospection of white objects. As a matter of fact, rather the opposite seems to be true: it is only by reflecting upon my external perceptions that I become aware of the internal states that accompany it. These three objections throw doubt on the claim that superstrong qualia are constitutive of perception in a way that weak qualia cannot emulate. Churchland’s weak qualia can do exactly the same work, but at considerably less philosophical expense.

In addition to the negative evidence against their superstrong competitors, there is positive evidence for weak qualia. Weak plasticity offers a principled explanation of several important aspects of qualitative consciousness. In the first place, it explains how qualitatively identical sensations may lead to different perceptions, and how qualitatively different sensations may lead to the same perception. Consider the following example (see figure 5.1). Suppose I submit myself to experimental surgery, and have my eyes replaced with infrared-sensitive organs of the kind described above. The rest of my physiology remains intact. After some training, I am able to see as well as to feel the temperature of objects. Although the kinetic energy detected by my
sensors is the same in both cases, it is intuitively plausible that the accompanying sensations will be qualitatively different. Weak plasticity respects this intuition, and even has a causal mechanism to explain it. By hypothesis, the different detectors in the example are connected to the same monitoring device. Qualitative differences of content can now be systematically explained as differences between detector states. A similar explanation can be given for cases in which identical contents lead to different perceptions. Thus, in Churchland’s original example, essentially the same detector was connected to different transducers (normal and infrared retinas), as well as to different monitoring systems. This situation is illustrated in figure 5.2. The detector state determines qualitative content, while the embedding system determines the semantic content of perception and the meaning of corresponding observation reports.5

By giving up the notion of a ‘proper’ content for qualia, the problem of arbitrary of disjunctive content is evaporated. Weak qualia do not need an

---

**Figure 5.1: Different weak qualia may lead to identical perceptions**

States of neural detectors $D_i$ are represented as vectors $(x_1, x_2, ..., x_n)$ in $n$-dimensional state space, as discussed in the previous chapter. These vectors are transformed into representations of perceived properties of the distal object by a neural net $T$, which functions as a monitoring device. Different detector states, read off by the same monitor, may lead to the perception of identical distal properties. The tactile or bodily detector for temperature (bottom), and the visual detector that is normally used for perception of color (top), are both used for perception of heat (kinetic energy). Arguably, the qualitative content of the detection states is different in both cases; by reference to the situation in ordinary human beings, the content of the visual detector is called ‘white’ in the diagram.
‘intrinsic’ specification, as they are simply irrelevant to the semantics of perception. Their qualitative content can be ostensively defined as whatever you feel when you are in detection state \(\{x_1, x_2, \ldots, x_i\}\). In addition, and rather surprisingly, the weak theory is also a vindication of some of the properties with which qualia are traditionally vested in philosophical analysis, namely, subjectivity, ineffability, possibility of inversion, and possibility of intersubjective empathy (cf. Dennett 1990; Churchland 1988a, 38-42). Weak qualia are radically subjective and tied to a unique point of view, because they are states of unique physiological detectors in individual observers. This also explains why we have no access to qualia other than our own, though considerations of analogy lead me to believe that your qualia will resemble mine. This analogy finds its neural basis in the shared physiology of observers: the more your neural vector space is different from mine, the harder it will be for me to imagine what your qualia are like. Moreover, the possibility of interper-

---

**Figure 5.2: Identical weak qualia may lead to different perceptions**

Identical detector states, read off by different monitors, may lead to perception of different distal properties. The same detector that is normally used for perception of color (triplets of electromagnetic radiation \(E_1, E_2, E_3\); top), may also be used for perception of heat (kinetic energy; bottom). Notice that D is connected to different transducers (different retinas) in the two situations. The qualitative content of the detection states is arguably the same in both cases; by reference to the situation in ordinary human beings, this content is ostensively identified as ‘white’ in the diagram.
sonal and intrapersonal inversion of qualia can be explained in terms of the inversion of neural vector space; no differences in behavioral effect will occur if the monitoring device is simultaneously changed in the same fashion (cf. Dennett 1991, 389ff). Finally, weak qualia are ineffable because they have no conceptual content of their own; they can be conceptually ‘exploited’ in indefinitely many different ways.  

3. Color qualia and the labeling fallacy

We saw that weak qualia can successfully break the dilemma of superstrength and superweakness of sensational content, provided they are construed reductively. Moreover, they can make an important contribution to our understanding of phenomenal consciousness. As intimated earlier, after a period of hesitant eliminativism cum epiphenomenalism, Churchland’s attitude toward qualia has taken a decisively reductionist turn (see, for example, his 1985a; 1986; 1988a, 28ff, 73ff, 146ff). In this section, I take a closer look at a characteristic example of the proposed reduction of sensations, namely, the case of color perception. After a brief sketch of how the notion of vector coding, introduced in chapter four, can be applied to the study of sensations, I consider the criticism that has recently been leveled against it by Fodor and Lepore, with particular regard to the example of chromatic qualia (1992, 187-207).

An example: chromatic qualia

It has been shown that our retinal cones, responsible for the detection of colors, consist of three populations of different cells, each maximally responsive to electromagnetic radiation in different bands of the visible spectrum. As illustrated in figure 5.3, the different pigments in retinal cones show characteristic absorption maxima for light energy at approximately 445 nm (blue light), 525 nm (green light), and 555 nm (red light). We also know that the color of objects is determined by the composition of the electromagnetic radiation reflected or emitted from their surface. Together, these facts make it highly plausible to surmise that, somewhere along the optical tract, the observed color of surfaces is neurally represented in function of the activation level of the three different types of retinal cones.  

Using the notion of vector coding discussed in the previous chapter, we can now represent each color sensation as a vector \( \langle x_1, x_2, x_3 \rangle \) in three-dimensional state space. By hypothesis, each component \( x_j \) tracks the activa-
The diagram shows the nominal shape of the absorption spectra of pigments in retinal cones. Three populations of cones can be distinguished, with maximum absorption values at ca. 450 nm (blue light), 525 nm (green light), and 555 nm (red light). For more details, see, for example, Fischler and Firschein 1987, 233ff.

The model of color perception fits in well with my discussion of weak qualia in the previous section. It exemplifies the way in which specific neural detectors may be identified, and the way in which their discrete activation states may be identified with specific sensations. It also illustrates the combinatorial virtues of vector coding. Ten discrete values along each of the axes allow the model to represent no less than 1000 different colors. (To appreciate the size of this number, recall that most people can identify between 150 and 200 colors.) Other sensory modalities, such as gustatory or auditory state space, may arguably be taken to wield more than three axes, each of which may take more values, thus enabling these systems to represent exponentially more distinct sensations. Thirty values along seven axes, for example, can harbor more than 20 billion distinct sensory states, each in principle tokening a qualitatively unique sensation.
The model has some additional virtues as well. Thus, it offers a reasonable explanation of our subjective sense of similarity between colors. Similarity corresponds to proximity in state space: the further apart their partitions are, the less likely two colors are to be counted as similar in phenomenal consciousness. Moreover, the model is able to account for several forms of color-blindness, including dichromacy (red-green color-blindness) and forms of anomalous trichromacy. If one of the three connections between eye and detector is impaired, or is lacking, then one dimension of color discrimination will fall away, causing the cube to collapse onto a plane. Subjects will typically display a diminished ability to discriminate colors along the missing dimension, as predicted by the model (Paul Churchland 1986; 1988a, 148). Alternatively, one dimension may be ‘dislocated’ as compared to that of normal subjects, causing subjects to systematically perform anomalously in color-matching tasks with regard to that dimension (anomalous trichromacy; Gregory 1987, q.v.).

Analogous to the work on color vision, various other elements of perception are now beginning to be understood from a neuroepistemic point of view.
view. Subjects of investigation include the perception of smell, taste, contour, curvature, depth, texture, pitch, melody, motion, balance, force, as well as a number of even more specific perceptions such as that of water surfaces, speech, human song, or faces (for an overview of classic and recent work on the ‘groundwork of cognition’, see Richards 1988; Harnad 1987). In all of these studies, a key role is played by the representational format used for encoding the information conveyed by the relevant transducers, analogous to the vector approach described above. In principle, each of these studies thus has access to an explanation of sensations and their qualitative contents in terms of detection states, according to the general approach described in this chapter, bringing sensations of all sorts and modalities within the reach of neural epistemics.

The labeling fallacy

In a recent critique of Churchland’s theory of perceptual contents, Fodor and Lepore (1992) have argued that the approach of vector coding is fundamentally mistaken from a cognitive point of view. Although their objections are directed in particular against the example of color perception, they obviously affect the status of neural representations as such.

The argument of Fodor and Lepore can be summarized as follows (op. cit., 187-207). The sensory state, \( \langle x_1, x_2, x_3 \rangle \), to which color sensations are claimed to be reducible, is in and by itself only a physiological state. Whatever qualitative identity it may have derives entirely from the dimensions of the state space in which it finds itself. The question, then, is how these dimensions are to be specified. If they are specified in terms of psychophysics (for example, in terms of the activation levels of specific populations of retinal cells), then the vectors do not represent qualia at all: the theory is simply not about color perception, although it may be an admirable exercise in the field of eye physiology. In the terminology introduced in this chapter, the alleged neural qualia are liable to become superweak. On the other hand, if the dimensions themselves are specified directly in terms of colors (blue, green, and red), then the theory may indeed be able to explain why a given vector represents a sensation of, say, brown (as a specific mixture of blue, green, and red), but it will ultimately be circular, because it simply presupposes the qualitative identity of vectors \( \langle 100, 0, 0 \rangle, \langle 0, 100, 0 \rangle, \) and \( \langle 0, 0, 100 \rangle \). These basic vectors are apparently assumed to have content intrinsically, which makes them superstrong, as I have called it. As Fodor and Lepore put it,
“What Churchland has is a dilemma: it may be that he isn’t intending to require that the dimensions of his state space correspond to properties of the contents of the mental states (objects, events) they taxonomize. In that case, he isn’t doing semantics at all. He’s doing, as it might be, psychophysics (…). If, on the other hand, Churchland is taking the talk about neural representation seriously, his move to state spaces leaves all the old problems about content identity still to be solved” (Fodor and Lepore 1992, 204-205).

According to Fodor and Lepore, the mistake in Churchland’s position is caused by the fact that he is unable to choose between elimination and reduction of qualia. As an eliminative materialist, Churchland should content himself with superweak qualia. “An eliminativist doesn’t need a notion of semantic identity. An eliminativist doesn’t want to reconstruct semantic discourse; he wants to change the topic” (loc. cit.). This, however, does not appear to be Churchland’s intention. What he wants is a notion of neural representations, and he is therefore obliged to account for their meaning. Hence, the hypothetical neural states must be invested with a proper content of their own—“for a semantics taxonomizes mental states by their contents, not by their causes” (op. cit., 200-201).

The gist of Fodor and Lepore’s objection against Churchland’s theory of color perception is that it commits the labeling fallacy, which we have already met in the previous chapter. The theory is accused of conflating the labels on the axes of state space with the states themselves. By assumption, the labels specify the qualitative content of the corresponding detector states. But these labels are visible only to the researcher, whereas they cannot be seen ‘from within’ by the detector itself; there is no intrinsic connection between content and detection state. In the previous chapter, I presented a general argument to the effect that the assumption of intrinsic content for mental representations is unwarranted. Now, with regard to weak qualia, the requirement is assuredly too strong. We saw that weak qualia can plausibly combine causal inertia with cognitive relevancy. They allow us to explain qualitative differences between states of consciousness as causal differences between detection states. No proper epistemic content is needed. Put in terms of the labeling fallacy, we can properly say that the labels worn by weak qualia are simply blank.

Probing deeper into Fodor and Lepore’s objection, we may discern an apprehension that the equation of sensations and detection states might
somehow be arbitrary: if axes do not determine the real meaning, then partitions can be said to represent anything at all. This fear is completely uncalled for, however. The qualitative content of weak sensations can be defined by ostensive means, complemented by a general understanding of the functional organization of the sensory system, and perhaps by considerations of analogy. This gives weak qualia all the ‘content’ they need. More content would be a burden instead of an asset: it would land us with the absurdities of ‘pure’, ‘self-identifying’ data of consciousness, whose ‘intrinsic contents’ are ‘interpreted’ by the subject in indefinite ways—the extravaganza of superstrong qualia.

4. Strong plasticity of perception

The apparent outrage of strong plasticity

The principle of strong plasticity entails that our sensations themselves may vary in function of a change of theory. New theories are accompanied by a qualitatively new perceptual consciousness. Theoretical innovations in the field of natural science, and in particular in the field of neuroscience, may cause our consciousness of self and world to expand in dramatic and unforeseen ways. In this section, I will examine three questions. First, what is strong plasticity? Second, does it commit us to superstrong qualia? And, finally, why are strong qualia to be preferred to weak?

In his paper on Churchland’s theory of knowledge, quoted earlier in this chapter, Philipse gives the following comprehensive statement of the principle of strong plasticity.

“According to the strong plasticity thesis, learning neurophysiological theory and being trained to use the sentence ‘The bipolar cells in my retinae are firing in the pattern XYZ’ on all occasions of our seeing a cat, will enable us literally to see or otherwise perceive these firing patterns on these very same occasions” (Philipse 1990, 133; see also op. cit., 132 and 171-172, n. 15).

If this statement correctly describes the thesis of strong plasticity, it is hard to find a theory that is more patently false. Barring supernatural intervention, no amount of training will ever enable me to “see or otherwise experience” Abraham Lincoln when I meet Mrs. Clinton, although I may of course be conditioned to use the spontaneous and non-inferential observation sentence
Qualia and epistemic content

‘President Lincoln’ every time I see Mrs. Clinton.

To some, the above description will suffice to demonstrate the absurdity of strong plasticity. Yet, I think it is rather the description that is at fault here. Let me mention three points on which the principle of strong plasticity has arguably been misunderstood. In the first place, notice that the examples used above apply the principle of plasticity highly selectively. They suggest that, *ceteris paribus* (leaving intact in particular all other linguistic dispositions), learning a single, isolated new observation sentence will yield a qualitatively new perception. All behavioral dispositions displayed by me indicate that I am perfectly able to tell Mrs. Clinton from Abraham Lincoln, to distinguish women from men, first ladies from presidents, and so on, with the sole exception of that one odd mistake. When asked to interpret my behavior, everyone will agree that I perceive Mrs. Clinton on all occasions of seeing Mrs. Clinton, in spite of the fact that I am adamant in referring to her as ‘Abraham Lincoln’. This has nothing to do with strong plasticity.

A second remarkable aspect of Philipse’s example is the fact that it replaces external perception with introspection. Thus, the proposal at hand is that we might learn to use a neurophysiological report as an observation sentence on all occasions of seeing a cat, which in turn would cause us to perceive our internal neurophysiology instead of the external cat. The external perception of cats is replaced by a qualitatively different internal perception. This bizarre result is admittedly not ruled out by strong plasticity, but it is not required by it either. The same effect can be produced equally well without the help of plasticity. Suppose that I decide, in a vaguely Buddhistic mood, to turn away from all worldly affairs in a most radical manner: henceforth, I devote my attention solely to introspection. Cats I no longer perceive. Perhaps, on one of my mental meanderings, a feline quale (as ordinary people prefer to call it) will cross my path, but I remain blissfully ignorant of its distal causes. In this example, introspection of internal states has come to replace my perception of external objects, yet without relying in the slightest on any principle of plasticity. Strong plasticity does not dictate that external perception be replaced by introspection, but it is perfectly able to explain why, after the replacement has taken place, I no longer perceive cats. By hypothesis, strong sensations change in function of our theoretical understanding of their causes. If the feline theory is replaced by an introspective one, my sensational consciousness will be limited to the proximal causes in the sensory detector, as opposed to their distal antecedents.

Finally, Philipse comments that no one “will seriously think that any
training whatever of our skill of visual perception will enable us to see or otherwise perceive the firings of the bipolar cells in our retinas” (1990, 132). Gentle mockery has apparently mollified acuity here. If true, the claim is irrelevant; if relevant, it is probably false, and dubious at best. If the claim is taken to be about the scope of visual perception, or about one of the other modes of external perception, in which we are asked to obtain a certain dexterity, it is certainly true. I will never be able to directly see or otherwise externally perceive the physiological processes going on inside my eye. But that is not at issue here. Strong plasticity entails the possibility of learning to perceive the physiology of the nervous system in introspective awareness. Thus understood, Philipse’s claim looses much of its plausibility. We can be taught to perceive many aspects of our bodily condition, as every sportsperson or physical therapist will readily acknowledge. Why should it not be equally possible to become aware of specific physiological states of our nervous system? “It must be a dull man indeed whose appetite will not be whet”, as Churchland remarks of the possibilities opened up by this projected expansion of self-consciousness (1979, 7).

How many strong qualia to one red apple?

These remarks may suffice to show that the initial description of strong plasticity is misguided in important respects. Yet, there is also some truth in it. Virtually all the explanatory work in my commentary was done by a distinction between introspection and exterospection. But how do these two compare to one another, and how are they related to their mutual qualia? To examine these questions, an analogy may be instructive. Arguably, introspection is not essentially different from external perception. To drive this point home, Churchland compares the brain’s multimodal measuring device with an ordinary amperemeter. The meter’s dial is calibrated to represent the strength of the electric current in the external circuit. Exactly the same instrument, if it is equipped with an appropriate division in gauss, can be used to measure the strength of the electromagnetic field inside the meter itself. The instrument is now operating “in introspective mode”, as Churchland graphically expresses himself (1979, 40).

Let me now try to apply this analogy to the example of color perception, discussed above. Figure 5.5 (see page 146) shows two observers, O₁ and O₂, who are both looking at an apple. If the apple’s surface reflects a% blue light, b% green light, and c% red light, then, crudely put, a neural vector \((a, b, c)\) will be activated in the chromatic state space of the observers. Both O₁ and O₂ are
equipped with a theory about the distal causes of their color sensations. This theory acts as the ‘dial’ of perception: it suberves the ‘conceptual exploitation’ of the observer’s sensations. Suppose $O_1$ has internalized the theory of reflectancy triplets, which he uses to spontaneously and non-inferentially produce the observation sentence, ‘The apple is reflecting light of composition $(a, b, c)$’. $O_2$ is conceptually retarded at this point; his observation report is simply that the apple is red. In addition to these external perceptions, both observers are also involved in introspection of their own detection states when looking at the apple. Here they use a different dial, that is, a theory of the internal organization of the human mind. $O_1$, who is using folk psychology, reports that he is having a sensation of red. $O_2$, by contrast, avails himself of a sophisticated theory of the neurophysiology of chromatic perception, which he has learned to use spontaneously and non-inferentially in his introspection reports. His report reads, ‘I am tokening a vector $\langle a, b, c \rangle$ in chromatic state space’.

How many different qualia are there in the example, and how are they related to one another? Recall that, if qualia were weak, there would be but a single sensation: the sensation of red, neurally instantiated as a vector $\langle a, b, c \rangle$ in color state space. This single weak sensation would then be conceptually exploited in four different ways. If qualia are strong, however, there seem to be no less than four different qualia, one for each theory. By assumption, each new theory may create its own, qualitatively different, sensational contents, based on the same detection state. At first blush, the consequences of this view are highly improbable. Our intuitions are strongly in favor of the view that there is only a single sensation involved, whether we choose to call it a sensation of red or one of $\langle a, b, c \rangle$. Moreover, this intuition is borne out by everyday experience: if we learn a theory about color, our qualia are not noticeably changed. Finally, the impending proliferation of strong qualia contradicts Churchland’s original proposal of reducing sensations to detection states: identical ‘causal middle-men’ produce four different qualia. The only apparent remedy seems to be to identify strong sensations with detection states plus theory—a solution that would make qualia conceptual, hence superstrong.

I think these conclusions are too rash, however. Strong plasticity entails that new theories may create new sensations, but not that this will necessarily be the case. If strong qualia remain unchanged in spite of a change of theory, then an additional explanation is required. I think that a good explanation can be given if we hold on to the original idea that sensations are reducible to detection states. We have defined weak qualia as ostensively identified detection states, leaving open the question of whether they are unique vectors...
or rather partitions (bundles of possible vectors) in state space. Now, I want to argue that the notion of strong qualia can be used to make this definition more precise. If this suggestion is correct, strong qualia turn out to be essentially a refinement of weak qualia, and not something entirely new to replace them. I propose the following set of interrelated definitions.

5. Two theories $T_1$ and $T_2$ conceptually exploiting the same vector space $V$ have identical strong qualia iff they use the same partitioning on $V$.

6. The strong qualia used by $T_1$ are different from those of $T_2$ iff $T_1$’s partitions differ from those of $T_2$. In particular, this will be the case if the partitions of $T_1$ are subpartitions of those of $T_2$, or conversely.

7. The perceptual grid of $T_1$ is different from that of $T_2$ iff it enables the subject to different discriminatory behavior, and different observation sentences in particular.

Figure 5.5: Proliferation of strong qualia?

Detection states $\langle a, b, c \rangle$ are fed into monitors for internal and external perception. To the left (see above), $O_1$ is using reflectancy theory for her external perception, and folk psychology for internal perception. To the right (see opposite page), $O_2$ combines standard resources for external perception with neurophysiology for introspection. The content of detector states, indicated by a question mark, is difficult to specify. Strong sensations are apparently proliferated by their close connection to theoretical exploitation.
For convenience of exposition, let me assume here that the perceptual grid of $T_1$ is a refinement of that of $T_2$. In other words, for each partition of $T_2$ there are one or more partitions of $T_1$, such that the latter are subdivisions of the first.

Let us now go back to figure 5.5, and compare the external perceptions of the two observers. $O_1$ has a sensation of red, while $O_2$ has a sensation of $\langle a, b, c \rangle$. Speaking completely generally, four possibilities suggest themselves. (A similar analysis can be given for internal perceptions.)

8. The ‘red’-quale and the ‘$\langle a, b, c \rangle$’-quale are identical, and $O_1$ and $O_2$ have identical behavioral dispositions.
9. The ‘red’-quale and the ‘$\langle a, b, c \rangle$’-quale are identical, yet the behavioral dispositions of $O_1$ and $O_2$ are different from one another.
10. The ‘red’-quale is different from the ‘$\langle a, b, c \rangle$’-quale, yet the behavioral dispositions of $O_1$ and $O_2$ are identical.
11. The ‘red’-quale is different from the ‘$\langle a, b, c \rangle$’-quale, and also the behavioral dispositions of $O_1$ are different from those of $O_2$.

The first of these possibilities, (8), is intuitively highly probable. In point of fact, it captures the original intuition behind the notion of weak qualia. Not-
withstanding the fact that O₂’s theory obviously contains the conceptual resources to make very finely grained distinctions, because the variables in descriptions such as \((a, b, c)\) may, in principle, take continuous values, these distinctions do not show up in O₂’s spontaneous phenomenal consciousness, although they may figure in his theoretical reports. This is evident from O₂’s overt behavior, and from his observation sentences in particular. Crudely but comprehensively put, there is no evidence that O₂ is better able to distinguish \((a, b, c)\) from \((a+\Delta a, b+\Delta b, c+\Delta c)\) than O₁ is.

Possibilities (9) and (10) are ruled out by the principle of strong plasticity as I have explained it here. Notice, incidentally, that (10), but not (9), is ruled out by weak qualia as well. Finally, proposition (11) represents the truly new possibility opened up by strong qualia—that the same detectors may be used in qualitatively new ways. How plausible is this suggestion? To illustrate this point, let me borrow an example from the ardent anti-physicalist Frank Jackson. The following passage from Jackson’s famous paper on qualia (1982) introduces Fred, an ordinary human being with one exceptional ability: he can distinguish two kinds of red where we can see only one.

“To him (Fred, JS) red₁ and red₂ are as different from each other and all the other colors as yellow is from blue. And his discriminatory behavior bears this out: he sorts red₁ from red₂ tomatoes with the greatest ease in a wide variety of circumstances. Moreover, an investigation of the physiological basis of Fred’s exceptional ability reveals that Fred’s optical system is able to separate out two groups of wavelengths in the red spectrum as sharply as we are able to sort out yellow from blue. I think that we should admit that Fred can see, really see, at least one more color than we can; red₁ is a different color from red₂. We are to Fred as a totally red-green color-blind person is to us” (Jackson 1982 (1990, 470)).

In this imaginary case, Fred can see one more color than we can; or, more to the point, he lacks one color but gets two in return. Our intuition tells us that Fred’s two kinds of ‘red’ sensations will be different from one another as well as from ours. This intuition is supported by the notion of strong qualia. What is more, strong qualia can explain the differences involved: Fred’s neural partition for red splits into two subpartitions, tokening of which results in sensations of red₁ or red₂. Fred was given by nature what Churchland claims can be given by theory, namely, qualitatively new sensations. The important point is the same in both cases: observers must comply with the boundary condition
stated in definition (7), above; that is, they must bear evidence of behavioral dispositions for making the relevant new discriminations. If this condition is fulfilled, the causes of these behavioral differences can be plausibly identified with strong qualia.\(^9\)

5. Theory and observation

Quality and epistemic content

The above discussion leaves us with two questions that are still open, namely, whether strong plasticity leads to superstrong qualia, and why strong qualia should be preferred to weak. The questions are related, as becomes apparent if we bring them into the form of a dilemma. If strong qualia are not essentially different from weak, then we do not need them at all. As we saw, weak qualia are a powerful enough tool in their own right. Yet, the only way for strong qualia to distinguish themselves from weak qualia is by becoming superstrong, namely, by forging a close alliance with theoretical exploitation.

The suggestion that strong qualia are really superstrong is fueled in particular by definitions (5) and (6), above, which entail that different theories may be based on different sensory partitions that are, by assumption, identical with different strong qualia. Hence, the theory’s conceptual grid is mirrored by the partitioning on state space. But apparently this can only mean that strong qualia are covertly endowed with a semantic identity of their own. Fred’s sensation of red\(_1\), for example, is a separate partition on state space, which acts as the necessary and sufficient cause of the corresponding observation report of seeing red\(_1\). All Fred’s internalized ‘color theory’ has to do, is simply to ‘read off’ the sensory partition that is activated in state space, and connect its message to the properly conceptualized observation report. Hence, the meaning of observation sentences will be determined by sensational content after all.

A serious shortcoming of this line of reasoning is that it underestimates the abstract nature of graphical expressions such as ‘state space’ and ‘partitioning’. In chapter four, it was explained that partitions are sets of possible activation states, that is, mathematically abstract representations of bundles of possible configurations of activity on a neural network. New strong qualia may involve a new taxonomy of detection states, but they are realized, strictly speaking, by the same old vectors and activation patterns. In this respect, strong qualia are not different from weak ones. The specific ways in which vectors are grouped
together in partitions is not determined ‘bottom-up’ by the vectors themselves, but is established ‘top-down’ by the theory exploiting them. There is no such thing as the ‘intrinsic’ partitioning of state space. If new strong qualia feel different from old qualia, as is intuitively clear from the example of Fred, then all we can say is that these differences are the result of different partitionings of detector state space, caused by the different ways in which theories process vectors. The different quality itself is identified ostensively, in the case of strong qualia as well as in that of weak. By the same token, strong qualia inherit all the virtues we found in weak qualia earlier in this chapter: they vindicate the traditional properties of subjectivity, empathy, inversion, and ineffability, and offer plausible explanations for differences and similarities in behavioral capacity across observers.

Summarizing the above, we may say that sensations do indeed possess qualitative identity, phenomenally present to consciousness, but that it does not follow from this that they also carry epistemic content of their own, which is subsequently interpreted by the observer. It is not so much the quality that is interpreted, but rather the interpretation that is qualified.

**Theory and observation**

With its strong emphasis on the role of interpretation in perception, the defense of qualia proposed here bears obvious kinship to the doctrine that all observation is theoryladen (Hanson 1958; Feyerabend 1962; Churchland 1992a), and to so-called ‘New Look’ theories of perception (Gregory 1974; Rock 1983). In one respect, this may be seen as a benefit of the approach, which is reinforced by (and in its turn lends support to) an established line of thought in contemporary philosophy of science. Similarly, the approach is backed up by a corpus of psychological evidence in support of the claim that there is no such thing as ‘pure’ observation untinged by conceptual background: cognition properly, as opposed to mere sensory transduction and physiological processing, starts with the ‘conceptual exploitation’ of the information picked up by the senses. Seen from the perspective of neural epistemics, this is tantamount to saying that not the sensory states themselves are important for cognition, but only their partitioning in state space.

At the same time, however, the association with theoryladenness invites a number of classic objections against my reading of strong qualia. Two such objections seem particularly relevant here, namely, that the conflation of observation and theory is incoherent, and that it is liable to cut science loose from its empirical foundations. On the view outlined in this chapter, all the
cognitively interesting work is done by *interpretation*, that is, by theoretical exploitation of sensory information. We saw that even *observation* is essentially a matter of interpretation, most notably in the case of strong qualia. Now, from the classical empiricist point of view, this rapprochement between theory and observation is threatened with incoherence. Observations are traditionally understood as ‘given’, in the sense that they are not the product of interpretation but rather *that which is interpreted*; the interpreting itself is the work of theory. But if observation is really interpretation, then what is there left to be interpreted? The entire notion of perception is threatened with conceptual incoherence once the delicate balance between observation and theory is disturbed. Moreover, that which is interpreted must exert some cognitive influence on *how* it will be interpreted. If this influence is denied, the link between theory and reality is radically severed. Apparently, this is precisely what happens on the above account, where the qualitative data of consciousness are stripped of all epistemic content. This consequence is made explicit by the second objection. The notion of strong qualia tends to obfuscate the way in which sensory information may ever come to constrain their theoretical exploitation: sensory states are claimed to be ‘interpreted’ by theory, but they have no content of their own to constrain this interpretation. Hence, our knowledge of reality must be essentially a free creation of the mind. Turning a phrase of Richard Rorty, the world seems to be well lost to us: if not outright idealists, we certainly are no longer empiricists (Rorty 1972; cf. Sleutels and Corbey 1992).

My comments on these objections will be relatively brief here, as I will have more to say on this subject in later chapters. I think the objections can be answered; but I also think that a truly adequate reply calls for a drastic reorientation of our general outlook on knowledge, rather than for a more penetrating analysis of the micromechanism of cognition. Still, this micromechanism holds some interest of its own at this point. Throughout this chapter, I maintained a sharp distinction between two radically different kinds of devices, one functioning as a *detector*, and another one functioning as a *monitor* of detector activity. At first sight, this is just a reverberation of the traditional distinction between ‘pure’ observations and ‘pure’ theory: the detector supplies what is ‘given’ (qualia), while the monitor serves to ‘interpret’ these data. This impression must now be corrected, however. The distinction between monitor and detector is really an artifact of the form of presentation, adopted only for ease of reference and graphical illustration. As a matter of fact, the monitor in neural nets is more properly seen as an *integrating part* of the detector *itself.* We
saw in chapter four that perception is a process in which neural activity is pushed through banks of synapses laden with prestored vectors. The resident vectors of weight matrices represent our theoretical expectations of what the world is like, and determine how the input activity on the net will be handled computationally. Monitor and detector are abstract aspects of this single, self-same net: they stand to each other not as two separate devices, but as weight matrix to activation pattern, or as partitioning to vector space.

It was pointed out in chapter four that the partitioning of ‘detector state’ space is only a mathematical description of differences in computational efficacy. It is the weight matrix resting on a cluster of nodes that defines these computational differences, and not the detector state itself. If an incoming vector $V$ is assimilated to a vector $P$ in the prototype region of figure 4.10, for example, its computational effects, modulo the weight matrix, are in relevant respects indistinguishable from those of the prototype vector: in other words, the detector state is interpreted in terms of prototype $P$ by the monitor. This means that epistemic content is granted to $V$ by the resident theory, but not that $V$ had content of its own prior to this processing.

Speaking in a slightly more abstract sense, the state space of a layer of nodes in one net may even be seen as mathematically partitioned by the computational processing performed by a different net connected to the first. This ‘remote partitioning’, as we may call it, simply mirrors the more distal differences in computational efficacy between a net’s various activation states. Whether or not this way of analyzing state space is fruitful depends on the connections between the nets in question, and on the cognitive functions they subserve; it depends on what we want to know about the way the net performs in the overall hierarchy. Thus, it is perfectly possible that two separate nets are related as interpreter and interpreted, in much the same way as the relation between monitors and detectors described earlier. Arguably, any theory that can be built into a ‘detection’ net can also be realized in a separate module, specialized in the interpretation of another module’s activity. This situation is not essentially different from that in which monitor and detector are aspects of the same net, however. Suppose two interconnected, modular nets $D$ and $M$ are related as detector and monitor, respectively. $M$ performs the interpretation of the information available in $D$’s activation states; in this sense, $M$ and $D$ are related as theory and observation. With regard to $D$, considered in isolation, the question of theory and observation simply repeats itself, however. $D$’s states are partitioned not only by virtue of $D$’s external connection to $M$, but also by $D$’s internal connectivity structure. Obviously, the
processing inside D itself is prior to its conceptual exploitation by M; in this sense, we may perhaps say that it is ‘pretheoretical’ in nature. Strictly speaking, however, D’s internal connectivity matrix is as much a theory as that of M: the ‘preprocessing’ is itself theoretical, leaving us with the conclusion that all observation is theoryladen, even if some processing may be more theoretical than other. Again, we see that there can be no such thing as ‘pure’ observation.\(^{10}\)

In a way, I believe that the first objection mentioned above is correct: the distinction between observation and theory cannot be coherently maintained in neural epistemics. But this does not necessarily mean that the approach is wrong; it may also be that the distinction itself is mistaken. I want to suggest that we bite the bullet and obliterate the distinction between observation and theory in its classical form. The idea that something like ‘pure data of consciousness’ must lie at the basis of all empirical science is an illusion created by bad philosophy—the philosophy of superstrong qualia, endowed with intrinsic mental content. If perception must be ‘pure’ in order to count as observation, we can simply do without.\(^{11}\)

Of course, none of the above is meant to say that our knowledge of reality is not empirical, or that the sensory input does not constrain our theory of the world. On the contrary, I want to press the idea that the contribution of reality to cognition may be much more direct than is assumed on traditional empiricist accounts. Viewed from this perspective, the neuroepistemic approach may be able to combine the insights of New Look theories with those of their opponents, such as Gibson’s theory of ‘direct perception’ (Gibson 1979; cf. Churchland 1989a, 228-229). I shall have more to say on Gibson’s theory in the final chapter. Observation does not yield a semi-finished product, such as ‘pure percepts’ or ‘raw data’, from the interpretation of which we must labor to reconstruct the true nature of reality. Observation and theory are related much more intimately, as two aspects of a single process. Any attempt to emancipate one at the expense of the other is doomed to fail. As we have seen in this chapter, this applies in particular to the qualitative data of phenomenal consciousness: qualia may be strong, but they certainly are not superstrong.

Expanding our sense of self and world

I conclude this chapter with discussing a residual worry regarding Philipse’s description of strong plasticity, quoted at the beginning of this section. In a way, Philipse’s rendition may be taken to suggest that strong qualia, understood as neural partitions, derive their true semantic identity from an intro-
spective theory. After internalizing the theory of human neurophysiology, we may be able to identify our detection states on introspective grounds alone. Thus, we may become aware of the individual vectors tokened in neural state space, and learn to see the various ways in which this space can be partitioned. Equipped with this new and more fine-grained introspective grid, would it not be possible for us to use it to improve our external perception as well? By studying the structure of my superstrong feline qualia, I should be able to learn more about the nature of cats.

It is true that Churchland, for one, often speaks as though a better understanding of our inner lives might teach us more about the external world. As we saw in chapter three, he stresses the fact that our current self-understanding, folk psychology, lags behind our view of the rest of reality. If we were to replace folk psychology by a new self-image based on modern neuroscience, the result would be a dramatic expansion and refinement of self-consciousness, or so it is claimed.

“This more penetrating conceptual framework might even displace the commonsense framework as the vehicle of intersubjective description and spontaneous introspection. Just as a musician can learn to recognize the constitution of heard musical chords, after internalizing the general theory of their internal structure, so may we learn to recognize, introspectively, the $n$-dimensional constitution of our subjective sensory qualia, after having internalized the general theory of their internal structure” (Paul Churchland 1986 (1989a, 106)).

Now, if introspection reveals the true nature and internal structure of strong qualia, does it not also follow that they have a semantic identity of their own? For what else can this structure be but the partitioning of state space, that is, the very same structure that instantiates the exploiting theory’s conceptual grid? Consequently, introspection may reveal aspects of our knowledge of the external world that we had previously been unaware of: it may add to our knowledge of the world, as well as to that of ourselves.

If this is indeed Churchland’s intention, I strongly want to oppose it. Consider the following example. Suppose I am unable to tell cats from dogs, for some reason that has nothing to do with the proper functioning of my brain, or that of my sensory system. Perhaps the samples I have been shown were collected from rare varieties of canoid cats and cat-like dogs. Whatever the actual cause, I am stuck with the fact that canine vectors and feline vectors are
swepted together in a single partition of the relevant detector space in my nervous system. The question, now, is whether introspection could teach me how to repartition my vector space in separate volumes for cats and dogs, such that I would henceforth be able to tell them apart in external perception as well. The answer must be negative. No amount of neurophysiological introspection will ever come up with anything but ‘anonymous’ activation vectors. Vectors do not come semantically ‘prepartitioned’, neither in introspection nor in exterospection. Partitions are the result of conceptual exploitation, and derive their meaning from them alone. To be sure, my sophisticated introspective theory may enable me to subdivide my partition for ‘cat-or-dog’ into two or more volumes, say A and B, but it will be impossible for me to tell which volume is feline and which is canine on introspective grounds alone. Someone else may of course inform me that, in normal human beings, the A’s belong to dogs and the B’s to cats, but then I will just have been taught a new theory for external cat-and-dog perception (cf. Churchland 1985a (1989a, 64ff)). Similarly, I may learn on my own to make the distinction in external perception, and subsequently become aware of the distribution of cats and dogs over A and B. Either way, external perception takes transcendental precedence over introspection, an insight also pressed on different grounds by Kant in his ‘refutation of idealism’ (Kant 1781, B 274ff).

To drive this point home, consider the following idea proposed by Churchland in his paper on eliminative materialism (1981). Churchland suggests that a new system of neurophysiological ‘Übersätze’ and ‘übersätzenal attitudes’ may come to replace the familiar propositional attitudes of folk psychology. This new system is claimed to be a much more powerful tool for understanding our inner lives—a suggestion with which I strongly sympathize.

“According to the new theory, any declarative sentence to which a speaker would give confident assent is merely a one-dimensional projection—through the compound lens of Wernicke’s and Broca’s areas onto the idiosyncratic surface of a speaker’s language—a one dimensional projection of a four- or five-dimensional solid that is an element in his true kinematical state. (Recall the shadows on the wall of Plato’s cave.)” (Churchland 1981, 89).

Surely, the parenthetical reference to Plato’s cave cannot be taken seriously. No amount of introspection will ever be able to teach us more about the ‘real world’ outside. Similarly, no matter how thoroughly we study the projector,
it will never tell us what the slides are about. True, a system of ‘übersätzenal attitudes’ would enhance our knowledge of knowledge (our νοησις νοησεως, as it has been called in Aristotelian philosophy), but not our knowledge itself.\textsuperscript{13} The same point can be made in terms of the dual aspect theory of the propositional attitudes, introduced in chapter three. Neural epistemics may be able to improve our understanding of the reality of knowledge (natural aspect), but not of its contents (logical aspect).

\textit{Qualia weak and strong}

Strong qualia are not superstrong. Their appearance of identity is really an abstraction in retrospect. They are reducible to neural detection states, but considered in themselves they are causally inert. In all of these respects, strong qualia are exactly the same as weak qualia. The differences between weak and strong qualia are represented by propositions (9) and (11), above. In one of my earlier examples, infrared observers were credited with the same qualia as we have when we see white, in spite of the fact that their discriminatory behavior is obviously different from ours. This example of weak qualia is an instance of possibility (9), above. If qualia are strong, this possibility will be ruled out, however. I do not think this is a great loss; rather, it seems to me to be a plausible correction of an intuition that was feeble to start with (see this chapter, note 5). The second difference between strong qualia and weak is represented by proposition (11): the emergence of new qualia that make a difference to overt behavior is explained in terms of a repartitioning of detection state space. This point is also a correction of the original doctrine of weak qualia. While weak qualia retained a sense in which their qualitative identity was ‘given’, the new view argues that this identity is better seen as ‘made’. Thus, strong qualia move us one step further away from the idea that qualitative content is somehow intrinsically determined.
Theories of content

The past few chapters have been extremely critical of the notion of ‘super-strong’ or ‘intrinsic’ mental contents. I have argued that this notion leads to absurd consequences, and, in particular, that it is the main obstacle in overcoming the naturalistic fallacy in epistemic theory. Meanwhile, however, I have been arguing for a theory of natural epistemics that is computational at heart. This poses a serious problem, for the received view of computational theories of cognition rests on a notion of mental content that comes very close to being ‘intrinsically determined’ in the sense rejected here. Computationalism seems to require a level of natural symbols on which the laws of mental computation are operating, and in terms of which cognition can be understood. If computationalism is to be a viable project, so it is often argued, then these symbols must have ‘computational contents’ of their own, which are variously claimed to be specifiable in purely ‘naturalistic’, ‘syntactic’, ‘formal’, or generally ‘non-semantic’ terms. Hence, there is a transcendental argument in favor of intrinsically determined contents in cognitive science.

Technically speaking, the question is one of taxonomy: how are the mental states that figure in the laws of cognitive science to be type-individuated, such that their taxonomy captures precisely all cognitively relevant distinctions and generalizations? Because, ex hypothesi, the relevant states in cognition are representations, or content-bearing states, the question is at the same time one of content: which taxonomy of content best fits the purposes of cognitive science? In recent years, this angle on the problem has given rise to a new subdiscipline in the philosophy of mind, the so-called theory of content (Fodor 1987; 1990a; Loewer and Rey 1991a; Von Eckardt 1993, ch. 6).

The theory of content is in many ways a textbook example of scholastic philosophy—a bewildering variety of academic puzzles and technicalities, liberally spiced with objections and distinctions. Over the past ten years, mental content has been claimed to be wide, narrow, opaque, transparent,
distributed, local, calibrational, translational, computational, correlational, functional, syntactic, formal, ecological, strict, autonomous, teleological, and symbolic, to name only some of the better known species of content that have been distinguished. In this chapter and the next, I intend not so much to partake in this new scholasticism, but rather to try to identify the underlying tendencies, the general structure of the dialectic. I shall only occasionally indulge in discussing more strictly technical issues.

The present chapter introduces two competing approaches to the determination of content in computational philosophy, which I shall call ‘internalism’ and ‘externalism’. After discussing a number of arguments in favor of internalism, I proceed to argue that they exhibit a repeating pattern of failure, based on a common presumption that mental content is necessarily intrinsic. Far from being required by computational theories of cognition, or so I shall argue, this presumption is an unwarranted, and unwanted, addition to computationalism, and ultimately goes back to a Cartesian metaphysics of subjectivity. A detailed discussion of externalism will be postponed until chapter seven.

1. Punch cards and mental content

A punch card heuristic

Computationalism in cognitive science is generally understood as the claim that the semantics of cognition is somehow determined by its syntax (Fodor 1975; Newell and Simon 1976; Newell 1980; Dennett 1981a; Haugeland 1981b; Cummins 1989; Block 1990, 135). I think it is important to be as graphic as possible about this idea, so I want to ask the reader’s indulgence for the following very simple example. Instead of the usual Turing machine analogy, let me introduce a punch card heuristic for mental representations. Punch cards, like Turing machines, are now almost archaeological objects. Yet, they are a perfect illustration of all the relevant properties of ‘formal symbols’ operated on by ‘computational rules’.

Punch cards are pieces of paperboard, or some similar material, perforated according to a predetermined code. The punches can be read by a machine, for example by means of photo-electric cells. If the machine is properly equipped with a decoding system, it can also determine which information is encrypted in the pattern of punches. Reading the content of a card is basically a matter of sorting: the punch pattern tells the machine into which category or categories a card falls, and what further action needs to be taken.
Figure 6.1 illustrates a simple perceptual case. The perceptual image of a distal object, in this case a cat, is picked up by a monitor. The monitor in turn is hooked up to a card puncher, which transforms the input to a punch card representation, or *percept*. This card is passed on to a sorting device, where the punch pattern is matched against that of a set of prestored master cards, representing *concepts*. When the pattern on the percept card is found to match that of a given master card, the appropriate action routines are initiated, such as printing out the report ‘CAT’.

Computationalism claims that mental representations are essentially like a set of punch cards. The mind’s punch cards are instantiated presumably in the form of vectors on weight space and on activation space, for a neural net somewhere in the brain. As explained in chapter four, the input card is presumably a vector in the *activation* space of a neural net’s input and/or hidden layer, while master cards may be thought of as prestored ‘prototype’ vectors determined by the net’s *connectivity* matrix.

Notice that the distinction between input cards and master cards is only a metaphor for the distinction between perception and recognition; they need not correspond to any separate *entities* or *real parts* of the machine. Thus, it is not necessary that the machine in figure 6.1 be equipped with pieces of paperboard acting as master cards. It suffices that, upon receipt of a given input card, the machine will respond with the appropriate action routines;

![Figure 6.1: A punch card heuristic for computation and representation](image-url)

Mental representations (percepts and concepts) may be compared to punch cards. The processing of punch cards in the ‘conceptual system’ is essentially a matter of *sorting*. Input cards from the perceptual system are matched against a resident stock of master cards. Master cards determine which further computational action needs to be taken, such as issuing a printed report ‘CAT’. (Adapted from Cummins 1989, 37.)
no explicit intervening search-match-and-act sequence is needed. In other words, it suffices that the operation of the machine can be interpreted as if it were using master cards. This does not necessarily mean that master cards are not real, however—it just depends on what you want to know about them. ‘Master cards’ are a name for the (possibly complex) causes of systematic differences in discriminatory behavior of which the machine is capable. These causes are certainly real. Yet, they do not necessarily pick out locally available entities, such as stored pieces of paperboard. Rather, they are abstract descriptions of different, complex machine states, distinguished in terms of their different computational efficacy.

Probably most of the machines that actually worked with punch cards did not use separate master cards. Their decoding system was wired straight into the machine’s operating structure. This severely limits the range of possible applications for which the machine can be utilized: it is capable only of the narrow set of tasks defined by its internal structure, such as cat recognition in the example above, or, in a more realistic example, transferring a sum of money from one account to another. By separating master cards from internal structure, the machine is turned into a multipurpose device: a new set of master cards running on the same machine will define a new set of tasks of which the machine is capable. The difference is one of economy, not of computation. Instead of changing the machine itself, only the cards are changed; meanwhile, computational effects are the same. For the time being, I will retain the idea of master cards as a useful fiction.³

Let us now put the punch card heuristic to work in the field of content. By assumption, each card has a certain informational or cognitive content, in virtue of which the machine relates to its environment in a meaningful way. For the sake of argument, imagine that each card, in addition to its specific pattern of punches, bears a printed label on which this content is stated in plain English. The machine cannot read this information, of course, though we can. Still, according to computationalism, a cognitive machine behaves as if it could read its own labels. Hence, computationalism requires that differences in content be mirrored by differences in punch pattern.

At first blush, it may seem that it is the intrinsic punch pattern of each card that determines its content. This is not true, however. What matters is not the pattern as such, but rather the pattern as read by a particular machine. The distinction is easy to draw in terms of our heuristic. Consider two machines, $M_1$ and $M_2$, that use roughly the same type of punch cards. $M_1$ is used for cat recognition, while $M_2$ is used in your local bank for registering
cash deposits. Now, it is perfectly possible that your cat will cause exactly the same pattern to be tokened in $M_1$ that a 200 dollar deposit to your account would cause in $M_2$. Although the punch cards are the same, the machines will read them in entirely different ways. Hence, the intrinsic properties of punch cards do not determine, or insufficiently determine, how they will be handled computationally.

Notice that the opposite of the above example may hold as well. It is perfectly conceivable that a given machine will treat different punch cards as if they were the same. Punch patterns may be much richer than is required by the machine operating on them. Some of a card’s perforations may simply not be recognized by the machine, or they may be ignored computationally. Either way, cards that differ only in irrelevant perforations will be treated as being identical.

What matters in punch cards is not what they are, but how they may be used. It is advisable to have some definite terminology for distinguishing the different types of properties involved. Let me introduce the following definitions, based in part on the distinctions drawn by Devitt (1990 and 1991).

1. A property $P$ of representation $R$ is subcomputational $=_{\text{def}}$ $R$ has $P$ in virtue of the specific physical substrate in which it is realized.
2. A property $P$ of representation $R$ is formal $=_{\text{def}}$ $R$ has $P$ solely in virtue of factors residing in $R$ itself.
3. A property $P$ of representation $R$ is syntactic $=_{\text{def}}$ $R$ has $P$ in virtue of its function in the overall system of representations of which it forms part.
4. A property $P$ of representation $R$ is semantic $=_{\text{def}}$ $P$ determines or expresses the cognitive content of $R$.

In the punch card heuristic, paperboard, weight, and physical size are typical examples of subcomputational properties. Formal properties include the number and distribution of punches an each card. The syntactic properties of punch cards advert to the rules operating on them, that is, to the specific ways in which cards are processed by a given machine. Finally, the semantic properties of punch cards are made explicit on their hypothetical labels, which state a card’s content. These distinctions are easy to draw in terms of our heuristic, but they may prove extremely subtle and difficult to draw in the case of natural cognition.
Master cards and mental content

The punch card heuristic highlights two important facts about computational theories of cognition. In the first place, computational theories seek to explain cognition specifically in terms of the *syntactic* properties of mental punch cards. Neither subcomputational properties nor formal properties will do: they may constrain the computational possibilities of a given card, but they do not explain which of them will be actualized by a given computational architecture. This result complies with some of the points made in earlier chapters. Thus, in chapter three I argued against eliminative materialism that the subcomputational study of neural wetware is simply unable to address the proper explananda of cognition; yet, pace functionalism, it may *constrain* the study of cognition in important respects. Furthermore, in chapters four and five I took issue with certain interpretations of connectionist modeling that consider only the intrinsic properties of activation patterns on neural nets. Against these, I argued that no merely formal account of vector spaces can explain cognition. Only vectors *in action* can, that is, vectors considered from the perspective of what neural nets can *do* with them.

The second important fact about computational theories is that, in order to count as an explanation of cognition, they must explain how computational systems are able to go *from content to content* in a meaningful way. In the punch card analogy, the machine must behave as if it were actually able to read the *labels* on the cards it is processing. By definition, however, the

---

**Figure 6.2: Properties of representations**

The diagram illustrates the relationship between various types of properties of representations as defined by computationalism (see propositions 1-4). Overlapping boxes at different levels indicate that the corresponding taxonomies of representations are presumed to coincide, or to be supervenient at least. Some important corrections to this diagram will be introduced below.
machine can only ‘read’ syntactic properties. Hence, computationalism requires that syntactically identical cards wear identical labels, and that different labels are worn by syntactically different cards (see figure 6.2). Semantics is determined by syntax, or, as Daniel Dennett has put it, the computational mind is a “syntactic engine driving a semantic engine” (Dennett 1981a; cf. Fodor 1975; Newell and Simon 1976; Newell 1980; Haugeland 1981; Block 1990; Von Eckardt 1993, ch. 6).

Where does all this leave the notion of ‘intrinsic’ content? On the above account, mental content is certainly not intrinsic in the sense that it would be determined by formal properties of the representation itself. Rather, content is determined by syntactic properties, the definition of which advert to the processing capacities of the system as a whole. Yet, the question will simply repeat itself with regard to the syntactic properties: for how are they to be determined? As it turns out, there is a general presumption in classical computationalism that syntactic properties are identified purely in terms of the internal structure of cognitive systems. This means that content must be intrinsic after all, namely, inasmuch as it is intrinsically fixed by the system’s computational proclivities.

The connection between classical computationalism and intrinsic content is particularly clear if we look at the notion of ‘master cards’. As explained above, master cards denote the complex sets of causes that determine how all other cards will be handled computationally; they may or may not be instantiated as separate entities. Master cards determine which formal properties of punch cards will be treated as computationally relevant, which will be ignored, and what their computational effects will be. They are just another name for a system’s intrinsic syntax. Now, with regard to master cards the question of content simply repeats itself: where does their content come from? The classical answer is that it does not come from anywhere. By hypothesis, computationalism is committed to the claim that content supervenes on whatever determines the system’s computational structure; hence, it must supervene on master cards. Master cards lend ‘syntactic content’ to all other cards, but they themselves do not derive content from anything else. Hence, their content must be intrinsic.

According to classical computationalism, the computational mind is defined by the set of master cards running on its ‘universal’ hardware. This interpretation of computationalism is known by many names, including ‘Strong AI’ (Searle 1980), the ‘Syntactic Theory of Mind’ (STM, Stich 1983), ‘High Church Computationalism’ (HCC, Dennett 1991b), ‘Good Old-Fash-
ioned AI’ (GOFAI, Haugeland 1985), or the ‘Computational Theory of Cognition’ (CTC, Cummins 1989).

Master cards challenged

In earlier chapters it was argued that any aspiring theory of cognition must start from folk psychology’s identification of the *explananda*. As we have seen, it is in particular the *content* of mental representations that is identified by folk psychology, namely, as the content of the propositional attitudes attributed to a subject. This requirement poses a serious problem for computational theories. As it turns out, the folk psychological taxonomy of content does not seem to match the syntactic taxonomy used by master cards. Hence, master cards may not be able to explain cognition.

The standard argument is due to Hilary Putnam (1975b). Putnam invited us to imagine a planet that is exactly like Earth, including all its inhabitants and culture. Twin-Earth, as we may call it, is an exact physical replica of Earth down to the molecular level. There is one difference, however: in place of water, Twin-Earth has another substance with chemical structure XYZ, which looks and behaves just like $\text{H}_2\text{O}$. In particular, XYZ and $\text{H}_2\text{O}$ are assumed to be phenomenologically indistinguishable. Finally, assume that chemical theory has not yet been developed on Earth and Twin-Earth. Now, consider Mary and her Twin-Earth counterpart, Mary*. By assumption, Mary* is an exact physical replica of Mary (with the exception of a considerable amount of XYZ). Because Mary and Mary* are assumed to have type-identical nervous systems, also their computational systems must be the same. Which mental *contents* Mary and Mary* are tokening when they are said to believe that water is wet, or when they express a desire for a glass of water? Arguably, the *content* or *aboutness* of these beliefs and desires is different on Earth and on Twin-Earth: Mary’s beliefs are about $\text{H}_2\text{O}$, whereas Mary*’s beliefs are about XYZ. Therefore, the contents of mental states are not determined, or are insufficiently determined, by the syntactic properties of computational systems. As Putnam put it, “Cut the pie any way you like, meanings just ain’t in the head” (Putnam 1975b, 227).

Putnam’s thought experiment is by no means uncontroversial. For one thing, it is based on dubious doctrines about natural kinds, possible worlds, and rigid designators (Kripke 1972; cf. Lakoff 1987). Moreover, the example is framed from the vantage point of an external observer, whose superior knowledge of the *real* denotation of ‘water’ is absolutely essential. The Twin-Earth case thus rests on a form of metaphysical realism, or ‘externalism’, as
Putnam has called it—a doctrine that has now officially been abandoned by Putnam himself (see, for example, his 1981 and 1987). It is not at all clear what the implications of this recantation are with regard to the question of content. (This is an issue to which I shall come back in the final chapter.) Finally, one may well wonder whether the example is really about folk psychology and computationalism. Folk psychology does not dictate that my beliefs about water be about $\text{H}_2\text{O}$ or XYZ. From the folk theoretical point of view, Putnam’s example would rather seem to be demonstrating the exact opposite, namely, that Mary and Mary* have the same propositional attitudes, for they are both said to believe that water is wet, refreshing, drinkable, and so on for the other phenomenological properties. Hence, Mary and her Twin are processing the same cognitive contents.$^5$

On the other hand, it is not difficult to devise other examples along the same lines, which are not flawed in the above respects, or at least less obviously so. For example, suppose Twin-Earth to be exactly like Earth, except that Mary’s hair is platina-blond while Mary*’s is raven-black. When queried, both Mary and Mary* will say their hair is blond. We readily attribute to Mary a belief that her hair is blond. But how about Mary*? Surely folk psychology dictates that we say that she knows her hair is black, but for some queer reason stubbornly refuses to call it that. The point can be made even more emphatically in terms of punch card machines. Imagine two identical machines to be involved in different tasks in entirely different environments. For example, one machine may be calculating the orbit of a recently launched Spaceshuttle, while the other is working on a chess problem. By some queer coincidence, the machines may be systematically tokening the same punch cards, master cards and all, and yet be processing entirely different contents.

Variations of the Twin-Earth conceit have been run by Putnam, Burge, and others to show that the contents of mental states are not determined solely by factors internal to the individual subject (see, for example, Putnam 1981; Burge 1979 and 1986; Stich 1978 and 1983). In all of these examples, the individual subject is kept constant, while certain other factors are allowed to vary. These other factors, which typically include environmental, sociolinguistic, cultural, and historical facts about the subject, are all intuitively relevant for the folk psychological determination of content. It seems to follow that all strictly ‘individualist’, ‘solipsist’, or ‘internalist’ approaches necessarily miss something vital for the determination of content.
Putnam's argument reversed

If Putnam is saying that the ladder is not long enough to reach the apple, Fodor says we should lower the tree. While Putnam argued that cognitive content, as we normally understand it, is not determined by intrinsic factors, Fodor concludes that the notion of content is at fault here. Obviously, the folk psychological notion of mental content is much *richer* than that of computational theory. Hence it must be attenuated to fit the purposes of computationalism. We must concentrate on what Twins have *in common*, instead of trying to make science of whatever differences folk psychology may attribute to them. In terms of our heuristic, if Twins are operating with the same master cards, and if they seem to be wearing different labels, then we should simply *rewrite the labels*.

The most radical conclusion from this line of argument is drawn by Stephen Stich (1983; 1991). His suggestion is that we forsake the notion of ‘content’ altogether. There is much to be said for this approach. If all that matters to cognition is determined by the intrinsic properties of master cards, then we can simply do without their labels. For, by assumption, the labels will only repeat what the master cards say; therefore, they can be missed.

“As I see it, the notion of ‘content’ or the folk psychological strategy of identifying a mental state by appeal to a ‘content sentence’, despite all its utility in the workaday business of dealing with our fellow creatures, is simply out of place when our goal is the construction of a scientific theory about the mechanisms underlying behavior” (Stich 1983, 5-6).

The problem with this type of argument for a ‘contentless’ psychology is that it cuts at the very root of cognitive science itself. Without content, no representation; but without representation, no cognition. As a matter of fact, as was noted in chapter three, Stich’s theory is a form of eliminativism. The only difference with Churchland’s brand of eliminativism is that the proper level of explanation, according to Stich, is not that of neuroscience, but that of the ‘autonomous syntax’ of cognitive systems. Yet, the position of Stich suffers from the same problems as Churchland’s. Its proposed elimination of folk psychology defeats its own purpose by eliminating the very explananda of cognitive science.

Acutely aware of this problem, other philosophers, most notably including Fodor, have drawn less radical conclusions, though not less extravagant ones. Over the past two decades, Fodor has argued for a version of compu-
tational theory that *retains* the notion of ‘content’, but that is otherwise virtually indistinguishable from Stich’s version. Basically, the idea is that Twins must be subsumed by the same *intentional* explanations, not just by the same neurological, biochemical, or syntactic explanations. This means that all differences in content that are *not* matched by ‘underlying’ differences in syntax must be irrelevant from a cognitive point of view. But this may still leave us with a notion of content, namely, with a sort of *narrow content*, defined as that part or aspect of folk psychological content that supervenes on intrinsic syntactic properties (see figure 6.3). Twin-Earth counterexamples to computationalism tend to focus on factors that form part of *wide content*. Wide content is by definition irrelevant to computational processing, as it is not determined by internal, syntactic factors.

*Two competing research paradigms: internalism and externalism*

The Putnam line and the Stich-Fodor line with regard to mental content are characteristic of two radically opposed approaches to cognition, which I shall

---

**Figure 6.3: Computationalism’s hierarchy for type-identifying representations**

The diagram shows a new and more sophisticated version of figure 6.2. At the semantic level, a distinction is made between ‘narrow content’ and ‘wide content’. Narrow content is defined as that part or aspect of a representation’s content that is determined by purely syntactic properties. It is intended to shield computational theories of cognition from Twin-Earth objections, and at the same time to avoid the problems facing Stich’s ‘contentless’ psychology. The computational level is sometimes called the ‘symbolic’ level, while the subcomputational level is also known as the ‘subsymbolic’ level. An additional new feature in the diagram is the tentative distinction at the subcomputational level. Notice the unfilled slot overlapping the formal properties at the computational level; this may be the place where, according to ‘subsymbolists’ such as Smolensky, or according to ‘implementationalists’ such as Fodor and Pylyshyn, the architectural properties of neural nets are located (see chapter four).
call ‘externalism’ and ‘internalism’, respectively. Different authors have used different names for the two paradigms, depending on the specific context in which the issue is raised. Thus, internalism is also known as ‘syntactic’ or ‘autonomous psychology’ (Stich 1983), ‘methodological solipsism’ (Fodor 1980), and as ‘individualism’ (Burge 1986). Similarly, the Putnam line has been called ‘naturalism’ (Fodor 1980; 1987; Stillings 1987a), ‘non-individualism’ (Burge 1986), ‘causal’ or ‘teleological semantics’ (Fodor 1990; Cummins 1989), and ‘wide functionalism’ (Kitcher 1985), to name only some of the most obvious candidates. The expressions ‘internalism’ and ‘externalism’ themselves, or some similar terms, are used by Colin McGinn and several other writers (see, for example, McGinn 1989; Heil 1992).

In most cases, the different terminologies reflect the fact that different aspects of internalism and externalism are highlighted by different arguments in their behalf. Yet, what the various species of internalism have in common strikes me as more important than what divides them; similarly for forms of externalism. Hence, my strategy is to bring out the shared assumptions first, and deal with the specific arguments later. For this purpose, let me define the following two gross positions.

5. Internalism =_{def} the individuation of explanatory kinds in cognitive science adverts only to factors that are internal to the individual subject.
6. Externalism =_{def} the individuation of explanatory kinds in cognitive science adverts in part to factors that are external to the individual subject.

Notice that these positions are mutually inconsistent, but only in a weak sense: strictly speaking, definition (6) admits the possibility that internal factors are relevant, although it rules out that they are exclusively relevant. In chapter eight I shall consider a redefinition of internalism and externalism that sharpens the contrast between them. Moreover, notice that the distinction between internalism and externalism in this context is different from Putnam’s homophonic distinction between two broad philosophical perspectives, also known as ‘internal realism’ and ‘metaphysical realism’ (Putnam 1981). There are, I think, important connections between these pairs of distinctions, to which I return in the final chapter.

If (5) and (6) define opposing positions in philosophy of cognitive science, the question suggests itself whether they define actual scientific practices, or normative scientific desirabilities. Stated in slightly different terms, are internalism and externalism put forward as descriptive, empirical claims, or as
normative requirements? It has been argued that highly regarded theories in computational psychology, such as Marr’s and Poggio’s account of early vision, make explicit reference to factors in the organism’s normal environment, thereby falsifying internalism. Hence, proposition (6) must be either a false description or a bad definition (see, for example, Kitcher 1985; Burge 1979; 1986; Stillings 1987a; Von Eckardt 1993, ch. 7; Garfield 1989, 147ff; Sterelny 1990, ch. 3). I do not think such descriptive arguments are very enlightening, however. For one thing, they may always be leveled by descriptive counterarguments on behalf of the opposite position, so that the score on this count remains even. Moreover, each purely descriptive argument can always be explained away as being irrelevant. Thus, if a theory violates proposition (5), it may be argued that it is simply not about cognition proper, but rather about (the computational side of) cognition plus something else. For example, Marr’s theory of stereoptics is arguably not (only) about the internal processing of visual information, but (also) about the fortuitous ecological constraints on stereoptics in mammals. The internalist may insist that cognition proper is the internal processing, not the environmental adventures of ambient light. Hence, what is cognitively relevant is Marr’s theory minus the latter, in compliance with proposition (5). Much the same strategy is open to the externalist. If, in defense of internalism, a descriptive argument is based on the intrinsic virtues of some connectionist model, or on the internal syntax of the ‘language organ’, for example, it may always be objected that something vital is left out, namely, an account of the content that is being processed by the internal machinery.

Descriptive arguments for internalism and externalism are useless unless they are combined with some independent normative argument. Without such additional support, they tend to get bogged down in a discussion of whether candidate theories of cognition are really about cognition, instead of being tuned at some other, ‘noncognitive’ level of analysis. Descriptive arguments evidently require a prior determination of what will count as an acceptable explanation of cognition to begin with. In earlier chapters, I have argued that cognitive science is committed to folk psychology as its a priori (but defiable) frame of reference. This will also be my position in the present context. Starting from the question of which notion of cognitive content is required by folk psychology, I will examine the various normative arguments put forward on behalf of internalism and externalism.

Figure 6.4 illustrates the difference between internalist and externalist approaches in a rather gross way. If we take a black box to be something the
contents of which can only be specified in terms of what comes in and what goes out, then externalism treats the subject as a black box, while internalism treats the world as one. According to internalists, what matters for cognition is only the information available within the bounds of your own nervous system; the world is whatever is beyond these bounds. Externalists, by contrast, adopt the opposite stance: the focal point of cognition is in the external world, outside the individual subject. Whatever happens inside the subject can be specified only in terms of its relation to the world. Notice that the diagram of figure 6.4 suggests that externalists are bound to treat the entire subject as a black box, in the fashion of behaviorism. This is an obvious exaggeration. Externalists typically advert to all the available information on the internal structure of individual subjects, no less so than internalists. Yet, the important point here is that the cognitive significance of this internal structure and internal processing, according to externalism, can be specified only in terms of the subject’s relations to the external world. The meaning of mental symbols is not determined inside the head.

Figure 6.4: Internalism and externalism. Black boxes or black worlds?
The diagram shows two fundamental options in cognitive research. Glass box/black world internalism (left) is also known as ‘autonomous psychology’, ‘syntactic theory’, ‘methodological solipsism’, or ‘psychological individualism’. It is the doctrine that computational theories of cognition should forswear the world beyond the individual subject, and study only the syntactic relations between internal symbols. Radically opposed to this is black box/glass world externalism (right), the most typical example of which is psychological behaviorism. Its central claim is that cognitive content is not determined by factors internal to the individual subject, but by the environment, including the subject’s broader sociolinguistic, cultural, and/or historical background. (Adapted from Dennett 1981a (1987, 53), who is quoting Keith Gunderson.)
The most striking aspect of internalism, from the point of view adopted in this study, is its claim that cognitive content is intrinsic. In the final analysis, content is determined by master cards, which must have content intrinsically. Externalism, as I have defined it here, is apparently not committed to this claim, or at least less obviously so. In the remainder of this chapter, I examine the case for internalism. A detailed analysis of externalist theories will be the subject of chapter seven.

2. Nine reasons for internalism

Varieties of internalism

The ur-version of internalism is better known as ‘methodological solipsism’. Originally devised by Carnap in the Logische Afbau der Welt (Carnap 1967, 64), the concept was redefined and introduced in the field of cognitive science by Putnam (1975b) and Fodor (1980). Methodological solipsism is a reverberation of Descartes’ claim that, even if reality were an illusion created by an evil demon, mental processes would carry on as usual. Hence, the relation between thoughts and the world is claimed to be irrelevant to the determination of their contents. The parallel with Descartes is explicit in both Putnam and Fodor. Putnam introduced it as follows.

“When traditional philosophers talked about psychological states (or ‘mental’ states), they made an assumption which we may call the assumption of methodological solipsism. This assumption is the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed. (In fact, the assumption was that no psychological state presupposes the existence of the subject’s body even: if \( P \) is a psychological state, properly so called, then it must be logically possible for a ‘disembodied mind’ to be in \( P \).) This assumption is pretty explicit in Descartes, but it is implicit in just about the whole of traditional philosophical psychology” (Putnam 1975b, 220).

Fodor used a number of different arguments to introduce internalism. Here is how he launched methodological solipsism in connection with computationalism, or, as he called it in this context, in connection with a ‘formality condition’. The formality condition, Fodor claimed,
“connects with the Cartesian claim that the character of mental processes is somehow independent of their environmental causes and effects. The point is that, so long as we are thinking of mental processes as purely computational, the bearing of environmental information upon such processes is exhausted by the formal character of whatever [is inside the head, JS] (...) I’m saying, in effect, that the formality condition, viewed in this context, is tantamount to a sort of methodological solipsism. If mental processes are formal, then they have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the semantic properties of such representations, including the property of being true, of having referents, or, indeed, the property of being representations of the environment” (Fodor 1980 (1981, 231), emphasis in original).

With all other forms of internalism, methodological solipsism shares the conviction that mental states are to be type-individuated by factors internal to the individual subject, irrespective of the subject’s embedding in a broader context. This shared conviction does not make internalism one definite theory, however. It is rather a many-pronged tendency, a collection of arguments centering around the same basic idea. This lack of definiteness makes internalism difficult to pinpoint, which is one of the reasons why the issue has engendered so much controversy in recent years. It is not incorrect to say that there are probably as many different forms of internalism as there are arguments to support it. In his doctoral thesis of 1984, Kevin Possin distinguished four or five principal lines of argumentation and as many distinct versions of internalism. Over the past ten years their number has more than doubled, and the identity of internalism is no clearer than before.

I have singled out nine arguments for scrutiny here. In order of appearance, I discuss:

- The argument from supervenience
- The argument from causal explanation
- The argument from autonomous behavioral descriptions
- The argument from neurophysiology
- The argument from functionalism
- The argument from computationalism
- The argument from conceptual role semantics
- The argument from scientific methodology
- The argument from epistemology
This list is not intended as an exhaustive enumeration of arguments in behalf of internalism; as a matter of fact, two additional arguments will be discussed in chapter eight, where they are raised as possible objections against externalism, to wit, the arguments from misrepresentation and from representational specificity. My purpose here is rather to establish a common pattern in all arguments of this kind—a pattern of failure to which all defenses of internalism fall prey.

1. The argument from supervenience

Many writers call on some form of physicalism for an explanation of the way in which mental content is determined by internal properties. The first four arguments all fall within this broad category. The principle of supervenience captures a range of very weak forms of physicalism (see, for example, Fodor 1974; Kim 1982; 1984; 1993; Haugeland 1982; Stich 1983, 164ff; Possin 1984, 92ff; Fodor 1987, ch. 2; Heil 1992, ch. 3; Von Eckardt 1993, 198ff; Papineau 1993, ch. 1). Informally speaking, one level of nature is said to supervene on another if the supervening level somehow depends on the more fundamental level. For example, the basic metabolic processes in physiology, such as respiration, digestion, and excretion, all supervene on biochemistry. The range of animal food is so great, however, that it is impossible to identify digestion with any one particular type of chemical process. At best, digestion can be ‘reduced’ to a (potentially infinite) disjunction of chemical types. Yet, it is generally acknowledged that the biological function of digestion has a biochemical explanation.

Put in slightly more formal terms, and applying the concept to psychological properties, supervenience says that there can be no difference in psychological properties without a corresponding difference in physical properties of the supervenience base. The reverse does not hold, however: not all differences at the base level are necessarily reflected by differences at the psychological level. Supervenience is stronger than mere token-physicalism or anomalous monism, yet weaker than type-physicalism or reductionism (see chapter two). Token-physicalism claims that all mental tokens are also physical tokens; no mention of differences between properties (hence, types) is made. Type-physicalism claims that each mental type can be identified with precisely one physical type. In between is the gray area of supervenience relations, which may take varying degrees of strength.

Two characteristic forms of supervenience have been distinguished by Jaegwon Kim (1982 and 1984). Consider two non-empty sets of properties $\Psi$
en $\Phi$ with elements $\psi$ and $\phi$, respectively. Paraphrasing Kim, we may now define a weak and a strong form of supervenience.

8. $\Psi$ is *weakly supervenient* on $\Phi = \text{def} \text{Nec } \forall x \forall y [\psi x \rightarrow \exists \phi \land \forall y (\phi y \rightarrow \psi y)]$

9. $\Psi$ is *strongly supervenient* on $\Phi = \text{def} \text{Nec } \forall x \forall y [\psi x \rightarrow \exists \phi \land \text{Nec } \forall y (\phi y \rightarrow \psi y)]$

Notice that weak as well as strong supervenience allow for the ‘multiple realization’ of mental states, discussed in chapter two. For any given property $\psi_i$ within $\Psi$, there can be two or more independent properties $\phi_1, \phi_2, \ldots$ from the base set $\Phi$ which serve as the supervenience base properties of $\psi$. The first (outermost) necessity operator in (8) and (9) ensures that, in every possible world, if something is $\psi$, then there will also be some associated $\Phi$ property $\phi_1, \phi_2, \ldots$. The second (innermost) necessity operator in proposition (9) makes the additional claim that, if a particular property $\phi_i$ is associated with $\psi$ in some possible world, it will be associated with $\psi$ in all possible worlds; $\phi_i$ is nomologically sufficient for $\psi$ to occur.

We may now complete the argument for internal content in the following way. There can be no difference in properties of mental states without a difference in properties of the supervenience base. In the case of mental states, the supervenience base is the physical make-up of the individual subject. Content is a property of mental states. Hence, differences in content must be determined by physical differences in the individual subject.

The argument from supervenience explicates, in effect, the internalist’s reading of Putnam’s Twin-Earth example: Twins must be tokening the same semantic types, because they share a common neurophysiological supervenience base. Stated in the above form the argument is fallacious, however. Supervenience, like Twin-Earth, may work in either of two ways, depending on the choice of the supervenience base. Externalists will typically claim that semantic relations supervene on causal relations between subject and environment; hence, they will opt for an external supervenience base. Internalists, by contrast, will typically claim that the supervenience base must be internal. It now becomes clear that the issue of internalism and externalism is logically distinct from that of supervenience. All arguments from supervenience presuppose that the choice between internalism and externalism has already been made. Hence, the above argument turns out to be circular. The crucial step is made in the second premise, where internalism with regard to the supervenience base is simply assumed without any further proof.
In addition to the objection of circularity, there is reason to expect that supervenience is more liable to work against internalism than in favor of it. ‘Representation’, ‘meaning’, ‘content’ and other such semantic notions, as tentatively defined by folk psychology, are essentially relational (or dyadic) predicates. This gives us prima facie reason to believe we should search for corresponding relational properties in the physical supervenience base as well. More succinctly put, cognition is typically a relation between subject and world; therefore, it may be expected to supervene on causal relations between the organism and its physical environment. Hence, as far as cognitive content is concerned, the argument from supervenience seems rather more comfortable with externalism than with internalism.

2. The argument from causal explanation

If it can be shown that physicalism in general is committed to a nonrelational taxonomy of properties, the case for internalism would be much stronger. We would have independent reason to assume that psychological properties supervene on nonrelational physical properties, that is, on local properties of the subject’s neurophysiology. This line of reasoning has been tried by Fodor (1987, 30ff). Fodor offers “a sort of metaphysical argument that individuation in science is always individualistic”, as he puts it. The notion of ‘individualism’, in Fodor’s sense of the word, is officially distinct from that of ‘internalism’ or ‘methodological solipsism’, as is explained by the following definitions:

“Methodological individualism is the doctrine that psychological states are individuated with respect to their causal powers. Methodological solipsism is the doctrine that psychological states are individuated without respect to their semantic evaluation” (Fodor 1987, 42).

Using these distinctions, Fodor offers a two-tiered argument in defense of internalism. The first stage of his argument leads from methodological individualism (MI) to nonrelational supervenience base, while the second stage leads from nonrelational supervenience to methodological solipsism (MS), that is, to ‘internalism’ as I have defined the term.

Fodor first argues that scientific taxonomy in general is nonrelational. According to the principle of MI, scientific properties are individuated with respect to their causal powers. This is a plausible assumption: only causally relevant similarities and differences can support the sort of nomological gener-
alizations used in scientific explanations. The next step is to show that MI is typically violated by relational properties. Fodor offers two arguments for this claim, one inductive, the other based on natural kinds. There are numerous examples of entities, properties, and processes in science whose causal powers are not affected in the least by relational factors. The causal powers of elementary particles, for example, are not affected by the changes of season, or by the fact that they were once part of the body of Baudelaire. Hence, there is inductive reason to believe that relational properties do not matter in science. The same point can be made in terms of natural kinds. We intuitively expect natural kinds to be specifiable by their ‘inner essence’, not by their relation to other things. Consider, for example, two broad categories of entities such as predators and objects the same age as Brigitte Bardot. The second category, a typical example of relational taxonomy, is not a natural kind: there are no laws of nature that apply to all and only the age cohort of Brigitte Bardot. Scientific taxonomy deals in natural kinds; hence, it is typically nonrelational. If this first conclusion is conceded, the second part of the argument leads to internalism along familiar lines, in the way explained above.

I have no quarrel with the principle of MI, as defined here: if things cannot make a causal difference, they cannot make a difference at all (for more discussion, see Von Eckardt 1993, 262ff). Yet, I do not think that MI, in spite of its misleading name, warrants the conclusion that individuation in science is generally nonrelational. In the first place, Fodor’s inductive argument is simply spurious; it is refuted by numerous counterexamples. In ethology, for example, fixed action patterns are standardly individuated in terms of environmental factors. In astronomy, the orbit of celestial bodies is determined in part by their intrinsic mass, partly by their distance to other celestial bodies, and partly by the intrinsic properties of these other bodies. In chemistry, the most salient causal properties of elements and compounds, such as their reaction characteristics, are specified almost by definition in relational terms; no mere consideration of an element’s ‘intrinsic’ place in the periodic table will suffice. Generally speaking, we find that relational taxonomies are the rule rather than the exception, in particular in the so-called ‘special sciences’. Hence, Fodor’s argument begs the question: it still owes an explanation of why relational properties should be avoided in the supervenience base of cognition.

The argument from natural kinds rests on a form of essentialism that is highly controversial (see, for example, Lakoff 1987). For present purposes, suffice it to notice that ‘natural kinds’ serve to explicate our intuitions about the way science works. Now, if science turns out to use relational taxonomies for
the entities, processes and properties to which its nomological generalizations advert, as I have just argued, should we not simply accept this as a fact about natural kinds? If intuition tells us that natural kinds are determined by their inner essence, then intuition must simply be corrected. Consider, for example, the category of predators, mentioned above. Fodor’s argument seems to suggest, in a most nearly Aristotelian way, that certain kinds of animals are predators because they have an inner essence which says, predator (or, in more modern terms, because they are equipped with a certain type of motor system, a certain type of digestive tract, and so on for the other facts about their internal structure that contribute to their being predators). Now, all predators doubtless have certain internal ‘causal powers’. Yet, these powers are not what is causally interesting about predators. Their being predators supervenes not on their internal, physical structure, but on their physical interaction with the environment: what matters are their foraging habits, the animals on which they prey, the structure of their ecological niche, and their position in the food chain. Succinctly put, the laws of predators do not advert to causes, but to effects.

The above conclusion is hardly surprising. What determines the choice of taxonomy in science is, generally speaking, not so much the internal ‘powers’ that may cause some effect, as the effects in terms of which these powers are identified. For purposes of explanation, it matters less what causes are (internal ‘powers’), than what they can do (powers ‘in action’, so to speak). This general result concerning scientific explanation harmonizes with some of my earlier remarks on the requirements of explanation in cognitive science. Thus, in chapter four I argued that aspiring theories of natural epistemics should concentrate not on the intrinsic properties of neural vectors, but on vectors in action, that is, on how cognitive systems can use them for systematically discriminating between different properties of distal objects. Hence, we reach the somewhat paradoxical conclusion that the principle of ‘methodological individualism’ supports anything but internalism with regard to mental content.

3. The argument from autonomous behavioral descriptions

If the second argument fails to rescue the first, perhaps a third argument can save both. Following up on my remarks about causal explanation, it may be suggested that, in cognitive science, the effects in terms of which internal causes are identified are themselves supervenient on internal physical structure. We saw just now that the taxonomy of internal causes presumably fol-
follows that of discriminatory behavior. This raises the question of how this behavior is to be type-identified. So far, I simply assumed that the taxonomy of behavior approximately follows that of folk psychology, and hence that it adverts to factors in the subject’s wider environment. These descriptions are obviously not supervenient on local internal structure. It has been argued, however, that folk psychological descriptions of behavior contain many elements that are spurious from a cognitive point of view. These unwanted elements must be eliminated to yield cognitively pure descriptions.

An argument along these lines is used by Stephen Stich to articulate his syntactic theory of mind (Stich 1983, ch. 8). Stich argues that folk psychological descriptions of behavior are infected with references to the historical, contextual, and ideological background of subject and observer. A notion of behavior that is thus relativized to a parochial perspective is ill-suited for purposes of scientific explanation. According to Stich, the explananda of cognitive science should be described in more objective terms. To this end, he introduces the notion of an autonomous description of behavior, defined as a description such that, “if it applies to an organism in a given setting, then it would also apply to any replica of the organism in that setting” (op. cit., 167).

“The thrust of the autonomy principle (…) is that by building historical, contextual, and ideological features into mental state descriptions, folk psychology has taxonomized states too narrowly, drawing distinctions which are unnecessary and cumbersome when we are seeking a systematic causal explanation of behavior. To believe that p is to be in an autonomous functional state and to have a certain history, context, and ideological relation to the ascriber. These further factors can surely be studied by various disciplines. But they have no place in a science aimed at explaining behavior. By slicing the pie too finely, they impede the formulation of those generalizations which apply equally to an organism and its replica” (op. cit., 170).

The ‘autonomy principle’ is proposed by Stich as a kind of conceptual filter for sifting the impurities out of folk psychology; what passes the filter is cognitively ‘pure’ behavior. The actual filtering is done by a replacement condition, which requires any ‘pure’ description of behavior to apply equally to a subject and to a replica in the same setting. The replacement condition superficially resembles Putnam’s Twin-Earth thought-experiment. Yet, it differs from it in inviting us not to imagine a physical replica in a different envi-
ronment, but a replica replacing the original in its *own* environment.

To give an impression of the kind of unwanted factors Stich has in mind, let me mention two of his examples. Consider first a welding robot on a factory production line. On a certain occasion, this robot is performing its millionth weld. Although ‘performing its millionth weld’ is a correct description of what the robot is doing, it is not an *autonomous* description. If, just prior to the weld, the robot had been replaced by a brand new physical replica, this new robot would perform exactly the same weld, but it would not be its millionth. The description, ‘performs its millionth weld’, makes reference to factors that no robot psychology can or should be held accountable for.

The second example is of a more human nature. Suppose you are interested in buying Stich’s car. After strained negotiations about price and delivery, you come to an agreement at last, and he sells you the car. Now consider the possibility that, just before the deal is closed, Stich is replaced instantaneously by a molecule-for-molecule replica. This replica will act in precisely the same way as Stich would have done in its place. It will sign the papers, count the money, and give you the keys to the car. *Stich* would thereby have transferred to you the ownership of the car. His replica, however, is necessarily unable to do so, since it does not *own* the car. ‘Transferring ownership’ is *not* an autonomous behavioral description. It adverts to a system of social, and more particularly, legal conventions for describing a subject’s behavior that need not correspond to any of the actual internal causes of behavior.

We now know something of the factors that should be *eliminated* from folk psychology’s description of behavior, but it still remains to be seen *what will be left*. What is autonomous behavior *itself*? This question is not raised by Stich himself, but, basically, two possible interpretations suggest themselves, one ‘eliminative’, the other ‘revisionary’ (cf. Stich 1983, 164ff; Von Eckardt 1993, 253ff). I want to argue that no argument for internalism can be gained on either of these interpretations.

The first and most radical way to read Stich is as restricting the factors relevant in behavioral descriptions to peripheral states of the actor’s body itself. To a rough approximation, autonomous behavior in this sense will consist of state transitions of the muscular system at the output side, and of the sensory system at the input side. Instead of describing a given action as ‘reaching for an apple’, for example, the autonomy principle will redescribe the action as a complex and coordinated movement of limbs.
Figure 6.5 shows some of the factors that may enter into behavioral descriptions. Folk psychology typically adverts to the leftmost and rightmost factors only, that is, to internal representations specified in terms of distal causes and effects, ignoring most of the intermediate processing going on between apples and apple bites; such is what I have called the ‘abstract’ nature of folk psychology. The radical interpretation of the autonomy principle, by

![Diagram showing the factors in behavioral descriptions.](image)
contrast, requires us to ignore the distal factors, and to concentrate on what is immediately on or under the skin: autonomous behavior is described in terms of proximal stimuli and proximal behavior, or, more radical still, in terms of peripheral (sensory and motor) responses. For present purposes, the distinction between proximal and peripheral factors can be ignored; the principal point of radical autonomy is that it restricts behavior to retinal, skeletal, or otherwise merely bodily events.

This first interpretation of the autonomy principle may properly be termed ‘eliminative’. When its conceptual filter is applied to folk psychology, just about nothing will pass the sieve. The descriptive taxonomy of folk psychology is based essentially on distal factors, and it is essentially silent about intermediate processing; if distal factors are sieved out, nothing remains. Radical autonomy is tantamount to replacing folk psychology by a descriptive taxonomy of an altogether different order. The eliminativist version of the autonomy principle is the position standardly attributed to Stich in the literature (see, for example, Fodor 1985 (1990a, 8ff); P.S. Churchland 1986, 395ff; Garfield 1988, 106ff; Sterelny 1990, 154ff). This interpretation is also supported by Stich himself, when he vents his emphatic pessimism about the future of folk psychology, and his sympathy for Churchland’s eliminative materialism (see, for example, Stich 1983, ch. 11; Ramsey, Stich, and Garon 1991).

Considered as an argument for internalism, eliminative autonomy faces two fatal problems. First, it begs the question of internalism, and secondly, it is in danger of eliminating the entire project of cognitive science itself. The problems are related, as becomes apparent if we bring them under the form of a dilemma: without eliminativism no argument for internalism, but with eliminativism no theory of content.

The first objection should be obvious: Stich does not argue for internalism, but simply stipulates that only factors supervenient on bodily states are relevant for scientific purposes. This is a mere statement of internalism, not an argument for it. Internalism is used for defending a particular notion of autonomous behavioral descriptions, and not the other way round. Stich himself is very plain about this. Thus, he writes in a characteristically prescriptive tone that the “basic idea of autonomy” is that

“the states and processes that ought to be of concern to the psychologist are those that supervene on the current, internal, physical state of the organism. (...) What this amounts to is the claim that any differences between organ-
isms which do not manifest themselves as differences in their current, internal, physical states ought to be ignored by a psychological theory” (Stich 1983, 164, emphasis added).

Stich is not interested in content; a fortiori, he is not interested in content determination. All he is interested in are, by his own definition, “current, internal, physical states”.

This brings me to my second objection: an eliminative reading of the autonomy principle defeats the purposes of cognitive science itself. Stich’s syntactic version of eliminativism suffers from exactly the same problems as Churchland’s neuroscientific version, discussed in chapter three. Folk psychology lays out the relatively observational vocabulary in terms of which cognitive phenomena are specified; to disregard it is to change the subject. Propositional attitudes specify the explananda of cognition, particularly as regards the content of cognitive states. If you are not interested in cognitive content, it is your good right to redefine your purposes in terms of neurophysiology, or in terms of ‘syntactic’ bodily events—but you will no longer be concerned with the explanation of specifically cognitive phenomena.

In previous chapters I urged the need for continual conceptual interaction between folk psychology’s descriptive resources and the developing explanatory apparatus of cognitive science. We saw that this opens up the possibility that folk psychology will be corrected and refined as science advances. A revisionary reading of Stich’s principle of autonomy respects this possibility. Recall that our question was how autonomous behavior is to be characterized. Against the eliminativist position, I argued that to act is more than to push your skin from the inside; it is systematic interaction with the environment, as specified, to a first approximation, by folk psychology. Autonomous behavior inherits from folk psychology this essential reference to distal factors, which it then goes on to correct and refine. It eradicates distinctions like those between ‘performing its millionth weld’ and simply ‘performing a weld’, or between ‘selling a car’ and ‘transferring ownership’, which are not relevant for purposes of explaining cognition. Hence, autonomous behavior redescribes the pertinent distal factors, but it does not make them any less external.

The attenuated version of autonomy suggested here violates internalism, yet it satisfies Stich’s replacement condition. Though referring to factors external to the individual subject, autonomous descriptions in the revisionary sense apply indiscriminately to a subject and to its replica in the same setting. Replacement by a replica rules out irrelevant descriptive factors, while sameness of
setting certifies that relevant external factors are kept in (cf. Von Eckardt 1993, 256ff).

4. The argument from neurophysiology

To wind up my review of physicalist arguments, I mention a familiar line of reasoning based on neuroscience (cf. Dennett 1981a; Churchland and Churchland 1983). The argument is, in effect, a species of eliminative materialism, discussed at length in chapter three. I confine myself here to some brief remarks.

If physicalism is correct, so one might argue, then whatever is of interest to cognition must be implemented by the nervous system. No account of cognition can be forthcoming that is not framed from this internalist perspective. Even if our ultimate aim is to understand how the brain enables the organism to cope with its complex environment, this account must be based on a prior understanding of the internal structure of the nervous system itself. Internalism is thus automatically built into the methodology of cognitive science. From the very outset, internal states are taxonomized in purely neural terms. Even when, at a much later stage, we consider the way in which the internal states hook up to the world, they necessarily retain their original type-identity. Therefore, neuroscience is committed to internalism.

The internalist sentiment in neuroscience is expressed in particularly concise terms by physiologist Horace Barlow. Consider, for example, the first of his five ‘neuronal dogmas’, which reads as follows.

“A description of the activity of a single nerve cell which is transmitted to and influences other nerve cells, and of a nerve cell’s response to such influences from other cells, is a complete enough description for functional understanding of the nervous system” (Barlow 1972, 380).

Barlow pleads for a pure and unmitigated bottom-up approach to cognitive phenomena. Starting from an understanding of the function of individual neurons, we may gradually ascend to understanding more complex neuronal ensembles, and eventually to understanding the function of the nervous system as a whole. The internalist taxonomy used by the approach is claimed to be ‘complete enough’ for cognitive purposes.

Essentially the same point was made by Paul Churchland in his defense of autonomous neuroscience, while similar sentiments underlie forms of connectionism that concentrate on the intrinsic structure of neural nets. I think these
views have been sufficiently dealt with in previous chapters. To summarize, let me quote Daniel Dennett on this issue:

“Psychology ‘reduced’ to neurophysiology in this fashion would not be psychology, for it would not be able to provide an explanation of the regularities it is psychology’s particular job to explain: the reliability with which ‘intelligent’ organisms can cope with their environments and thus prolong their lives. Psychology can, and should, work toward an account of the physiological foundations of psychological processes, not by eliminating psychological or intentional characterizations of those processes, but by exhibiting how the brain implements the intentionally characterized performance specifications of subpersonal theories” (Dennett 1981a (1987, 64)).

If mental states are studied from the internalist perspective of neural structure, no account of content or of cognition can be forthcoming. Any theory of neural epistemics is required to endorse folk psychology’s externalist taxonomy of content. This taxonomy may be refined and corrected, as we have seen, but it can neither be ignored nor eliminated.

5. The argument from functionalism

I now turn to a second family of arguments, based on functionalism rather than physicalism. The first of these is an argument that is rarely stated in explicit form, but that is never quite absent either. It is often found lingering under the surface of other functionalist defenses of internalism, which is why I shall simply call it ‘the argument from functionalism’. In many ways, it is the exact opposite of the argument from neurophysiology. While the latter attempted to prove internalism by tying mental states to their specific biological substrate, the present argument tries the same by severing the link between mind and body. Consider, for example, the following passage from Robert Cummins (1989), in which the author takes issue with Millikan’s externalist account of mental content. Computational theories of cognition (‘CTC’ in the quotation), Cummins argues, necessarily require an internalist taxonomy of mental content.

“The CTC goes farther than simple physicalism; it asserts that, in order to preserve the identity of a cognitive system (if not a whole mind or person), it suffices to produce a computational duplicate. Two systems running the same program on the same data structures are, according to the CTC,
For better or worse, the CTC seeks a theory of cognitive capacities of the sort that might be brought to bear on radically different environments (with differing success, no doubt), and that might be realized in radically different stuff” (op. cit., 81).

In this passage, two well-known aspects of functionalism are brought to bear on internalism, namely, multiple realizability and functional identity. The argument is largely implicit, but presumably something along the following lines is meant. Consider first functional identity. According to functionalism, the mind is an abstract constellation of functionally characterized capacities. Each of these capacities, and each of their component subcapacities, is type-identified in terms of its local i/o-control, and in terms of its place in the wider web of functions. Hence, functional structure is necessarily internal; it can be specified without reference to external factors. In the passage above, this point is made in terms of programs and data structures, which are claimed to be specifiable entirely by their internal functional structure.

Secondly, consider multiple realizability. According to functionalism, mind can be realized in radically different substrates. The level of explanation in cognitive science is logically distinct from that of biology and other such disciplines that study a particular kind of substrate. Now, if properties of the substrate are irrelevant for purposes of type-identifying mental states, then so will be the relations between substrate and physical environment. Hence, functionalism is committed to a purely internalist taxonomy of mental states.

The two lines of reasoning explicated here convey essentially the same thought, namely, that mind is a functionally self-contained entity, distinct from both body and world. If functionalism were indeed committed to this assumption, it would be a powerful argument in favor of internalism. However, I do not think the antecedent is true. A closer look at the notion of functional identity reveals that nothing is to stop us from widening the web of functional relations beyond the ‘virtual skin’ of the mind. Two different forms of functionalism can now be distinguished, as Patricia Kitcher (1984 and 1985) has argued: ‘narrow’ functionalism, which type-identifies mental states in purely internal terms, and, ‘wide’ functionalism, which identifies mental states in terms of their external functional relations as well. Notice that wide functionalism, despite its reference to external factors, is still a species of functionalism. Hence, we see that internalism is logically prior to functionalism. The argument from functionalism simply presupposes internalism, namely, by relying on the assumption that functionalism is necessarily narrow. This assumption, however, is false.
Functionalism as such is indifferent with regard to internalism and externalism. In addition, however, it may be argued that wide functionalism is to be preferred to its narrow alternative. Folk psychology type-identifies mental states by their content, the specification of which typically adverts to environmental factors. Functionalism in cognitive science follows the descriptive taxonomy of folk psychology. Hence, cognitive science is committed to a form of functionalism that accounts for external factors, that is, wide functionalism. From this perspective, the idea that narrow functional states hold some cognitive relevance of their own, determining some kind of ‘narrow’ content, becomes highly implausible.

The same conclusion applies to the alleged multiple realizability of mental states. The assumption that mind is a ‘disembodied’ entity, which should be studied in isolation from all bodily constraints, is simply gratuitous. It appears to be a relic of Cartesian metaphysics, which has somehow survived the revolt against dualism. Mind is not a free-floating substance, nor can it be studied as one. We have met several objections to this approach, in particular in chapters three and four. The exclusive top-down methodology inspired by radical functionalism suffers from a lack of specific constraints on the nature of cognition; by construing its explananda too abstractly, it is doomed to be vacuous. Cognitive science must advert to the specific purposes subserved by internal representations, that is, to the way representations are used by the organism to organize its interaction with the environment. Without body or environment, neither representation nor cognition remain.

6. The argument from computationalism

An argument of similarly Cartesian flavor may be derived from a consideration of computationalism. In Fodor’s classical defense of methodological solipsism, a so-called formality condition is formulated, which says that computational theories of cognition must type-identify mental states by their formal properties alone. As was already quoted at the beginning of this section,

“If mental processes are formal, then they have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the semantic properties of such representations, including the property of being true, of having referents, or, indeed, the property of being representations of the environment” (Fodor 1980 (1981, 231)).
Heeding the definitions introduced earlier in this chapter, the formality condition is more accurately described as a *syntactic* requirement. Computational theories must type-identify mental states by their computationally salient properties, that is, by properties to which computational processes are sensitive. These are, by definition, *syntactic* properties. Insofar as the system is processing representations, also the *content* of these representations must be specifiable in purely syntactic terms. Now, syntactic properties are taxonomized in terms of internal factors; therefore, computationalism transcendentally requires internalism.

To evaluate this argument, let me first reconstruct it in terms of the punch card heuristic. According to computationalism, the internal computational structure of a cognitive system is determined by its inner syntax, that is, by its set of master cards. Punch cards are processed by virtue of their syntactic properties (punch patterns), which cause the relevant master cards to be tokened. In addition to this, each card bears a label stating its representational content. The machine behaves *as if* it can read these labels, but in reality it has access only to punches. Now, suppose we were to admit labels, or other such references to external factors, in our explanation of cognitive processes. We would then effectively be positing homunculi inside the machine, for, by assumption, the machine itself can neither read the labels nor peek at the world beyond the punches. Homunculi *can*, but only by making the explanation circular. Therefore, on pain of circularity, cognitive science is committed to internalism.

The reconstructed version of the argument from computationalism makes it easier to see what is the crucial step in this line of reasoning: *we*, the explainers, must ignore external factors (labels) *because the machine does*. I think this inference is anything but clear. In particular, it is difficult to see how it is supposed to follow from computationalism without the aid of additional assumptions. Let me mention two such assumptions here, one methodological, the other of metaphysical.

The first assumption underlying the argument from computationalism is of a methodological nature: cognitive science must sharply distinguish between computation and interpretation—not the scientist’s *interpretation*, but the machine’s bare *computation* determines which cognitive content the machine is processing. As Patricia Kitcher put it in her criticism of methodological solipsism,
“Computational psychologists must not only acknowledge the gulf between theories of interpretation and theories of computation, they must limit themselves to the latter. Computational psychologists should forswear the world beyond the subject and consider only the formal relations among inner formulae” (Kitcher 1985, 88-89).

So long as we think of purely artificial systems like punch card machines, whose computational structure is given a priori, this assumption may seem rather trivial. If computational properties (punch patterns) are simply given, they are readily thought of as determining internal contents-for-the-machine, independent of our external interpretation of them (labels). As soon as we turn to systems whose computational structure is not given, however, the distinction between computation and interpretation becomes obfuscated. This happens most notably when we consider natural epistemic systems. In natural systems, computational structure is typically discovered (rather than given) by attending precisely to its relevant interpretation (rather than computation), that is, by analyzing its wider function in the system’s interaction with the environment. Consider, for example, the chromatic processor discussed in chapter six. The epistemic function of color perception, subserved by some neural structure N, is identified not by N’s internal structure, but by its use in the organism’s interaction with colored objects. Again, the corresponding vector space for color perception may reflect aspects of neural connectivity structure, but its computational structure is identified in terms of the wider function it subserves. Finally, if we discover that some partition on activation space represents ‘red’ rather than ‘blue’, we type-identify computational properties by their interpretation, and not the other way around.

Moreover, even assuming that we have complete knowledge of the brain’s internal computational structure, specified in purely ‘non-interpretational’ terms, all of this knowledge (although invaluable by itself) would still not amount to a theory of cognition. This objection is essentially the same as that against the argument from neurophysiology, discussed above. An analogy may serve to drive my point home. Suppose you ask me how to get from A to B by car. My answer will typically advert to the various possible topographical routes that take you from point of departure A to point of arrival B. In principle, each of my answers (each route) may also be couched in terms of the movements of the car’s internal parts: movements of the wheels, engine, gear, brakes, steering wheel, and so forth. Yet, there is no relevant description in terms of these internal movements that will generalize all and only the ways
to take a single route from A to B. Moreover, if there are several possible routes from A to B, there is no relevant generalization over all and only these possible routes in terms of the car's internal motions. The case is analogous in cognitive science. If you drive your mind from A to B in the internalist fashion, you will never know what you are doing. Similarly, if cognitive science explains your mental journey from A to B in terms of 'pure computations', it will have missed your journey’s quintessence. We may conclude that the methodological requirement underlying the argument for internalism is highly suspect: not only is it typically violated by computational theories of natural cognitive systems, but it also threatens to excavate the explanatory value of cognitive science as a whole.

At first sight, a further assumption in the argument from computationalism seems to obviate these difficulties. It is tempting to grant that interpretation is indeed indispensable for heuristic purposes, as I have just argued, but to insist that, metaphysically speaking, content is still determined from within. Regardless of the question what is given to us, the outside observers, the system itself has access to nothing beyond its own computational structure. Therefore, whatever content the system is processing must ultimately be understood from the internalist perspective.

I think this line of reasoning rests on a particularly pernicious prejudice in classical computationalism, which is identified most easily in terms of the basic metaphor on which it relies. Thus, we have seen that cognitive systems are taken to ‘have access’ to some factors but not to others. For example, they cannot ‘peek at the world beyond the senses’, or ‘read’ labels without the aid of homunculi. Again, cognitive systems are pictured as determining mental content ‘from the inside’, or as ‘interpreting’ content from the syntactically ‘given’. Similarly, it is claimed that computational structure may not be ‘given’ to us, the external observers, but that it is nonetheless all that is ‘given’ to the system itself. All of these expressions rely on a single master metaphor, namely, that of the Cartesian subject of cognition. The argument from computationalism urges us to take the point of view of this subject: locked inside the syntactic engine, it is denied access to the world beyond, condemned to read only the ‘computational content’ of its internal symbols. Only from this perspective does it make sense to think of cognition as the internal processing of intrinsic contents. This metaphysical assumption underlying the argument from computationalism, which inherits the worst of Cartesian metaphysics, should clearly be rejected.8

Summarizing the above, we find that computational theories of cognition
do not require that computation is divorced from interpretation, nor that cognitive systems are ‘really’ processing only syntactic symbols. Rather, computational epistemics makes it possible to address the question of how computational structure is used by cognitive systems to organize their interaction with reality. (Notice, though, that there is no hidden user inside the system; rather, the ‘user’ is simply the system itself.) Evidently, this question cannot be answered unless we interpret the computational activity of cognitive systems in terms of their relation to external factors. Hence, far from being an argument for internalism, computationalism is more readily combined with an externalist approach.

7. The argument from conceptual role semantics

A final argument based on functionalism calls attention to important features of mental content that apparently cannot be accounted for by external factors. Examples are phenomena such as referential opacity, semantic holism, conceptual change, epistemic holism, and various other ‘subjective’ properties of belief fixation. In addition to being rather ill-understood, these phenomena have in common that they have all been used to support some version of conceptual role semantics, which in turn is taken to be an argument for internalism.9

As an example, consider the phenomenon of referential opacity (Chisholm 1957; Fodor 1980; Block 1986). A well-known fact about intentional states is that their semantics tends to violate Leibniz’s Law. From the fact that Luke Skywalker believes that P(a), and the fact that $a = b$, it does not follow that Luke believes that P(b). For example, from the fact that Luke believes that Darth Vader must be killed, it does not follow that he believes that his father must be killed. He simply does not know that Darth Vader is his father. Generally speaking, the content of propositional attitudes appears to be represented by the subject under a description. Hence, there is no freedom of substitutivity for co-referring expressions. This feature is explained by conceptual role semantics, which claims that the specific ‘descriptive angle’ from which a symbol represents its object is determined by the symbol’s relations to other symbols, not to external factors.

Similar arguments for conceptual role semantics are fostered by semantic and epistemic holism, conceptual change, and other such high-level features of cognition. They all centre on the intuition that mental content is a question of how concepts hang together internally. Meaning is determined by conceptual schemes and by internal data structures, which may change even when the world does not. Hence, some version of conceptual role semantics must be
correct; therefore, internalism is correct.

On the surface of it, this line of reasoning looks suspiciously like an argument from despair: “some features of cognition are still ill-understood, so we may as well claim support for internalism from them”. Nobody really knows how to deal with the phenomena in question (for some state-of-the-art discussions, see Fodor and Lepore 1992). Strictly speaking, they are anomalies for internalism as much as for the externalist approach. For the sake of the argument, however, let us suppose that some version of conceptual role semantics is correct. The question, then, is whether conceptual role semantics is better explained by internalist approaches than by externalism. I think that, generally speaking, conceptual role semantics is no better off than functionalism in this respect. Applying the principle of methodological individualism, discussed above, we find that there can be no difference in conceptual role or ‘descriptive angle’ that is not a causal difference as well. The question of internalism and externalism now repeats itself with regard to these causal differences. Are they to be described as wide causal differences, or as narrow ones? Both options are still open. What is more, there is a general argument in favor of the wide option rather than the narrow one.

Take opacity as an example. The argument from conceptual role semantics suggests that the cognitive system itself constrains the interpretation of the symbols it is processing. Thus, symbol $S_1$ in the mental organization of Luke Skywalker would mean ‘must-be-killed’, and $S_2$ would mean ‘is-my-father’, because of their relation to other internal symbols. Now, presumably, Luke Skywalker is a computational system like you and me. This means that his descriptive angle on Darth Vader, tokening either $S_1$ or $S_2$, must make a specific computational difference. How to describe this computational difference itself—in terms of ‘narrow’ effects, or in terms of ‘wide’ differences? As we have seen, interpretation typically takes us beyond narrow computational effects. It requires us to take into consideration the wider causal functions subserved by computational structure. This is certainly true of contents such as ‘is-your-father’ and ‘must-be-killed’, which reflect important differences in a subject’s behavior as well as in that behavior’s environmental consequences. (Freud’s purported proof to the contrary never appealed to me, though it did to my father). Hence, the interpretation of mental symbols may be constrained by internal structure, but it is certainly not given by it.

The essentials of conceptual role semantics are at least as compatible with externalism as they are with internalism. An externalist example is Dretske’s account of the way in which mental symbols co-determine each other’s “selec-
tive sensitivity” to different aspects or “informational components” of the environment (Dretske 1981, 180ff). In a similar vein, Dan Lloyd has suggested a “dialectical” or “multiple channel condition” on the determination of mental content. On his view, what a symbol represents depends in part on its ability to synthesize the information carried by two or more prior symbols or input channels (Lloyd 1989, ch. 3). Finally, also the microfunctionalism espoused by Clark (1989, 34ff), and the neural epistemics outlined in chapter four above, assimilate relevant aspects of conceptual role semantics. In all of these cases, the frame of reference is strictly externalist. Therefore, conceptual role semantics cannot serve as an argument for internalism.

8. The argument from scientific methodology

One of the best-known arguments for internalism, due to Fodor (1980), is based on the claim that the alternative position, externalism or ‘naturalistic psychology’, as Fodor calls it, is a methodological impossibility. The argument may be summarized as follows. If a mental state has semantic properties, these are presumably fixed by the subject’s relation to his environment. Any scientific account of these relations must be based on nomological generalizations linking mental states with distal objects and their properties. Hence, externalists will need some “canonical way of referring to the latter” (op. cit., 249). In other words, the characterization of the objects of cognitive states must express nomologically necessary properties of these objects. This, however, means that externalists must wait for the results of natural science. If the object is salt, for example, the appropriate projectible characterization (‘NaCl’, or some successor notion) will be “available only after we’ve done our chemistry”.

“The theory which characterizes the objects of thought is the theory of everything; it’s all of science. Hence (...) naturalistic psychologists will inherit the Earth, but only after everybody else is finished with it. (...) No doubt it’s all right to have a research strategy that says ‘wait awhile’. But who wants to wait forever?” (op. cit., 248).

Fodor concludes that we must settle for internalism, which is the only way to avoid the methodological extravaganza of externalism.

The problem with the argument from methodology is that it is simply too fastidious. Fodor’s demand for no taxonomy without strict nomologicality puts a requirement on cognitive science that is not met by the rest of science
(cf. Stich 1983, 162ff; Kitcher 1985, 91; Garfield 1988, 62ff). If this demand were justified, no science would ever be possible at all. All scientific disciplines rely to some degree on the conceptual and descriptive resources of other disciplines. If they should all be waiting for each other, then none would ever get started. In actual scientific practice, no requirement of this sort can be found. The rule of thumb in science is reliability rather than nomologicality: any taxonomy is admissible as an inventory of a given domain of reality, provided there are no specific reasons for doubting its validity. As we have seen in chapter three, this is true in particular of the relatively observational resources of folk theories, which remain as a more or less constant frame of reference for the development of more sophisticated, scientific theories in their domain. If cognitive science is allowed to work by this rule, it should have free access to all the available information about reality in science and in common sense. In particular, it should avail itself of the taxonomic devices of folk psychology for type-identifying mental content.

9. The argument from epistemology

A possible reply to the above objection would be that the constraints on cognitive science are more severe than those on other sciences. Being heir to the traditional project of epistemology, cognitive science should also adopt its ancestor’s high standards of evaluation. In particular, it should be a foundational discipline, whose task it is to proof that our knowledge of self and world is justified (Rorty 1979). Obviously, a proof of this kind must be strictly a priori: if it relies on any empirical evidence itself, it will be viciously circular. Hence, cognitive science must abstract from all empirical evidence about reality. In particular, its central explanatory concept of mental content must be free from any such semantic notions as being true of objects in the environment. Therefore, mental content must be construed internally.

It is difficult to take Fodor’s argument from methodology quite serious unless something like the above argument is taken in as well. In point of fact, Fodor’s reasoning is a clear echo of Descartes’ method of universal doubt. So long as our description of the external world is not certifiably ‘canonical’, Fodor seems to argue, it should be doubted to the extreme: that is, it should be ignored for purposes of understanding cognition.

In chapter three, I pointed out that global doubt of the Cartesian kind will inevitably end in epistemic paralysis. Doubt is local by definition, in philosophy as well as in science. It takes certain beliefs as an Archimedean point for casting doubt on others. Descartes believed he had found a sort of ‘universal’
Archimedean point in the certainties of subjective consciousness, construed for this purpose as a mental substance locked outside the world. However, no such dualism is acceptable for epistemology today. The modern subject of cognition is anything but a non-physical Fremdkörper—it is an organism endowed with computational capacities, rooted firmly in its natural surroundings. This view of ourselves draws on evidence from many sources, including psychology, neuroscience, and computer science, as well as biology, physics, chemistry, and other sciences of self and world. Serving as the framework for understanding cognitive phenomena, this collection of evidence is presupposed by modern theories of knowledge, although this does not make it immune to revision: if specific reasons for doubting its validity come to light, it will be duly adjusted.

If we cease to think of ourselves as mental substances locked outside the world, then the individual subject is no longer a Cartesian vantage point from which to study cognition. Yet, the Cartesian temptation appears difficult to resist, even within a naturalized framework. Many scientists today seem to argue that the subject, though not locked outside the world, should still be locked inside the body. As psychologist Ulric Neisser once put it (before his conversion to Gibsonian externalism),

“There is certainly a real world of trees and people and cars and even books. (…) However, we have no direct, immediate access to the world, nor to any of its properties” (Neisser 1967, 3).

The “central problem of cognition”, as Neisser called it, is then how the subject manages to commute its ‘direct’ knowledge of what is inside to gain ‘indirect’ access to the world outside. I think this is simply wrong-headed. The physical organism is not a mere container for the subject of cognition; rather, it is the subject itself. Hence, there is no internal, Cartesian vantage point from which to study cognition. The subject’s point of view is that of a participant, not that of an internal outsider to reality.

The same point can be made by a consideration of the dual aspect theory of folk psychology, introduced in chapter three. There I argued that the propositional attitudes have a logical as well as a natural aspect. This also applies to the subject of cognition: it is real because it is a part of nature, and it is logical because it is identified as the ‘I’ in sentences such as ‘I believe that p’. Now, psychologists readily admit that the real ‘I’ forms part of natural reality, as an organism in its physical environment. Yet, many theorists appar-
ently hold on to the idea that the logical ‘I’ is somehow behind or inside the real ‘I’, where it has ‘direct access’ only to the real subject’s interior, and defines the subject’s internal ‘point of view’. This is clearly erroneous. The logical ‘I’ is not additional to the real ‘I’—they are two aspects of the self-same organism.

Far from being transcendentally required by the nature of epistemology, internalism seems to be the inferior position here. If epistemic theory is to explain how organisms use internal representations to reliably organize their interaction with reality, then internalism is badly inadequate. It can neither explain the organism’s use of representations, nor its interaction with reality, nor why this interaction is reliable.

3. A Cartesian heritage

A repeating pattern of failure

Gathering up the results of the foregoing discussion, we find that the arguments for internalism exhibit a repeating pattern. They all fail in similar ways, which can be summed up as follows.

8. *Circularity.* As independent arguments for internalism they fail because internalism is presupposed in each case. This result was to be expected: the arguments are typically framed as transcendental arguments; hence, they aim to establish internalism as a necessary precondition of certain important and highly valued requirements (physicalism, functionalism, explanation, etc.).

9. *Optionality.* They also fail as transcendental arguments, however, because internalism appears to be a merely optional position. By themselves, the above-mentioned requirements are logically independent of internalism. In particular, they are also compatible with externalism.

10. *Irrationality.* In addition, the arguments for internalism also fail as inferences to the best explanation, because they are all demonstrably more comfortable with externalism. Hence, the choice for internalism is irrational.

11. *Cartesianism.* In the final analysis, the case for internalism rests on a piece of fossile Cartesian metaphysics. Although Cartesianism is officially contradicted by modern psychology, parts of it survive in the presumption that the subject of cognition is inside the computational system, where it is processing internally defined mental contents.
These four points sufficiently establish the inadequacy of internalism as a research strategy in cognitive science. By the same token, they are prima facie evidence that some form of externalism is needed in cognitive science. Whether the forms of externalism proposed in recent literature are really a viable alternative to internalism, is a question I shall address in more detail in the next chapter. Before doing so, I want to take a final glance at internalism’s connection with intrinsic content. For the purpose of the present study, which is overcoming the naturalistic fallacy, this connection strikes me as what is most essential, and also most objectionable, in internalism.

**Internalism and the naturalistic fallacy**

Paradoxically, one way to see internalism in cognitive science is as an attempt to avoid committing the naturalistic fallacy. It may be argued that internalism, by stripping representations of their relation to the world, divorces the question of psychology from that of epistemology. This view of internalism is taken by Richard Rorty. In his crusade against mental ‘givens’ that intrinsically reflect the world as it is in itself, he explicitly exempts the Fodorian brand of cognitive science from his critique.

“Fodor’s picture of the mind as a system of inner representations has nothing to do with the image of the Mirror of Nature I have been criticizing. The crucial point is that there is no way to raise the skeptical question ‘How well do the subject’s internal representations represent reality?’ about Fodor’s ‘language of thought’” (Rorty 1979, 246; the reference is to Fodor 1975).

In other words, internalism cannot be accused of relying on intrinsic contents, because symbols in the language of thought may be *intrinsic*, but they are certainly not *contents* in the ordinary sense of the word, that is, they are not representations of the world. Succinctly put, cognitive science may be about ‘psychological’ knowledge, but it is emphatically not about ‘epistemic’ knowledge-of-the-world. This is what Rorty means when he claims that the ‘skeptical question’ cannot be raised with regard to cognitive science.

I think this analysis badly misjudges the effects of internalism on cognitive science. The ‘skeptical question’ *should* be raised! What is more, it can be raised as a truly *skeptical* question only if internalism is presupposed. And, finally, it can be *answered* only if it is *not* raised as a Cartesian question, and if internalism is rejected. Let me briefly explain these points. First, Rorty’s analysis
suggests that cognitive science should aim at knowledge that is not-of-this-world. However, this is simply not knowledge (which is why Rorty thinks cognitive science is not prone to the naturalistic fallacy). Knowledge is essentially of-the-world: cognitive science must explain how organisms use representations to reliably organize their interaction with reality. If it does not address this ‘skeptical question’, it simply fails as an account of knowledge. Secondly, the question of how well we represent reality takes a specifically skeptical turn only if it is assumed that we are subjects locked inside an organism. If we do not have ‘immediate access’ to the outside world, it becomes extremely doubtful indeed whether we can know it at all. But internalism is not the solution to this problem; it is rather its very cause. The final point should be clear now: only by abandoning internalism will cognitive science be able to address the proper explananda, and avoid the skeptical implications of Cartesian epistemology.

Eliminating content may be a cure for the naturalistic fallacy, but it is worse than the disease. Rorty’s analysis follows the false dilemma discussed in the previous chapter, namely, that all content must be either superstrong or superweak. If it is superstrong, it begs the question of cognitive content, while, if it is superweak, it is not cognitive. Choosing the second option, Rorty eliminates cognitive science as such. In addition to this, Rorty oversees the fact that the internalist reading of ‘superweak content’ is still an attempt to reduce knowledge to what is ‘given’ to an inner subject, namely, as symbols defined over internal syntax. Rorty’s analysis rejects the wrong thing about intrinsic content: in order to avoid the naturalistic fallacy, we should not reject content, but the idea that it is determined intrinsically.

Narrow content and the Fodor-Stich paradox

The same misunderstanding about intrinsic content emerges if we compare the most recent defenses of internalism. The state of the art in internalism today is represented by theories such as Block’s (1986), Fodor’s (1987; 1990b; 1990c), Stich’s (1983; 1991), and Cummins’s (1989). What binds these positions is a common allegiance to the idea that mental content is essentially two-tiered, consisting of a nucleus of narrow content surrounded by a cloud of additional constraints. Narrow content is presumed to be the proper object of computational theories of cognition; supervening on internal factors alone, it is determined intrinsically. Narrow content becomes intentional content if it is combined with certain additional factors, external to the computational system as such. These external factors are variously identified as ‘semantically interpret-
ing’ (Cummins), ‘contextually anchoring’ (Fodor), ‘causally and socially constraining’ (Block), or ‘culturally embedding’ (Stich) the putative nucleus of intrinsic content.

A typical example of this semantic dualism is Cummins’s account of mental content. According to Cummins, the computational theory of cognition (CTC) defines a notion of “simulation-based” mental representation (s-representation), such that all computational systems tokening the same s-representations will “simulate” the same semantically evaluable behavior, irrespective of their hardware, environment, or causal history. The actual intentional content of any given s-representation can then be derived from its s-content, provided

Figure 6.6: Two-tiered explanations of mental content
A nucleus of ‘narrow’ or ‘computational’ content is surrounded by a cloud of additional conditions which serve to (re-)establish the link with environmental truth conditions (left). Typically, nuclear content defines the taxonomy of explanatory kinds in cognitive science; the study of additional constraints is relegated to other sciences such as biology. (See, for example, Block 1986; Fodor 1987 and 1990a; Stich 1983 and 1991; Cummins 1989.) The diagram on the right shows the architectural companion to NC/AC semantics. In this scheme, the proper object of cognitive science is commonly claimed to be the universal central system, the ‘logical core’ of cognition. The study of the ‘human’ interface, including the sensor and motor channels, is then relegated to natural sciences such as biology. On the distinction between the central system and its ‘modular’ interface, see Fodor 1983, and chapter two, above. Well-known opponents of this type of architecture are Lakoff (1987) and Johnson (1987), who have argued that we should put the body (the human interface) back into the mind (the universal system).
“that we combine a causal or adaptational account of intentionality with the CTC, so that s-representation plus some further constraint external to the CTC (e.g., causal or adaptational or social context) equals intentionality: \( s\)-representational content + FC = intentional content” (Cummins 1989, 141).

Essentially the same position is taken by Fodor when he defines narrow content as a “function from context to truth conditions”: semantic evaluation follows when narrow content gets ‘anchored’ in context. Narrow content is what gets to be semantically evaluated given a context of interpretation (Fodor 1987, 44ff).

The logical consequence of this approach is that narrow content itself must be radically contentless. This conclusion is particularly clear in Stich’s view of narrow content. As we have seen, the strict reading of his principle of autonomy leads to a view on which content is simply eliminated. Fodor, too, is laudably explicit about this consequence. Unlike Stich, however, he thinks cognitive science can still remain a fully ‘representational’ or ‘intentionalist’ account of cognition.

“Narrow content is radically inexpressible, because it’s only content potentially; it’s what gets to be content when—and only when—it gets to be anchored. We can’t—to put it in a nutshell—say what Twin thoughts have in common. (...) One wants, above all, to avoid a sort of fallacy of subtraction: ‘Start with anchored content; take the anchoring conditions away, and you end up with a new sort of content, an unanchored content; a narrow content, as we say.’ (...) What is emerging here is, in a certain sense, a ‘no content’ account of narrow content; but it is nevertheless also a fully intentionalist account.” (Fodor 1987, 50–53, emphasis added)

This paradoxical result with regard to narrow content, which we may call the Fodor-Stich paradox, has puzzled many writers (see, for example, Stich 1983 and 1991; Cummins 1989, 124; Block 1991; Devitt 1991, 96; cf. Fodor 1991, 280ff). If cognitive science studies ‘narrow’ or ‘autonomous content’ (and not the additional constraints), then it studies representations insofar as they are contentless. From this shared assumption, Stich concludes that cognitive science does not study representations at all, while Fodor insists that it does. I think the proper way to deal with this paradox is to renounce the underlying assumption—it is simply not coherent to claim that mental content can be studied from a ‘narrow’ or internalist point of view.
My conclusion here was anticipated long ago by Leibniz in a famous passage of the *Monadology*. Leibniz suggested the following thought-experiment, a distant ancestor of Searle’s Chinese Room.

“Suppose that there were a machine so constructed as to produce thought, feeling, and perception, we could imagine it increased in size while retaining the same proportions, so that one could enter as one might a mill. On going inside we should only see the parts impinging upon one another; we should not see anything which would explain a perception. The explanation of perception must therefore be sought in a simple substance, and not in a compound or in a machine” (Leibniz, *Monadology*, par. 17. Quoted from: Leibniz 1934, 181; transl. Parkinson).

“We are obliged to confess”, so Leibniz concluded, “that perception and that which depends on it cannot be explained mechanically, that is to say by figures and motions”. To which we may add: not by mere figures and motions, as specified from an internalist point of view. For purposes of explaining cognition, what matters is not what is given to the internal spectator, but how it may be used by the machine as a whole. This, I take it, is how modern philosophy should construe the notion of “simple substance” referred to above.
Chapter seven
Externalism and mental representation

“A kingdom for a world

“O, what a world of unseen visions and heard silences, this insubstantial country of the mind! What ineffable essences, these touchless rememberings and unshowable reveries!” In these, the opening words of his study on consciousness, psychologist Julian Jaynes admirably paints some of the insidious seductions of the Cartesian mind.

“And the privacy of it all! A secret theater of speechless monologue and prevenient counsel, an invisible mansion of all moods, musings, and mysteries, an infinite resort of disappointments and discoveries. A whole kingdom where each of us reigns reclusively alone, questioning what we will, commanding what we can. A hidden hermitage where we may study out the troubled book of what we have done and yet may do. An introcosm that is more myself than anything I can find in a mirror” (Jaynes 1976, 1).

It is understandable that we feel reluctant to renounce the private kingdom of consciousness, the inner repose of desire and belief. Yet, this is precisely what the previous chapter urged us to do: to wake up from the Cartesian dream of subjectivity, at least insofar as it is tied to what Daniel Dennett (1991) has dubbed the ‘Myth of the Cartesian Theatre’, or what Gilbert Ryle described as the idea that the mind is lined with “phosphorescent” mental items, whose existence and nature are “inevitably betrayed to their owner” (Ryle 1949, 15). The doctrine to be opposed is that of ‘superstrong’ or ‘intrinsic’ mental contents, which are claimed to be immediately ‘given’ to the ‘internal’ subject.

In the previous chapter, I identified the internalist approach to mental content as a major source of philosophical confusion with regard to the problem of knowledge. In its stead, some form of externalism should be accepted, acting on the supposition that mental content is determined in part by factors external to the psychological subject. Contents are not coined in the imaginary kingdom of the mind, but derive in part from the world. The forfeit of the
Cartesian mind may not be such a bad thing after all, then: if we stand to lose a kingdom by giving up internalism, externalism may win us a world in its stead.

Internalism, in all its hues and varieties, is still the received view in computational psychology today, but the number of dissidents is growing. In the course of the past decade, externalism has begun to reach beyond the academic circle of mere philosophical dispute. Philosophers of mind are now making tentative contact with research programs in other sciences, such as ‘ecological’ or Gibsonian perception theory, evolutionary biology, and ethology. Views developed in these fields often corroborate the misgivings about internalist psychology discussed by philosophers. I shall come back to some of the implications of this rapprochement in this chapter and the next.

Although my sympathies lie more with externalism than with internalism, I have certain apprehensions about the type of externalist theories floated in recent literature, which I hope to make clear as we proceed. My plan in this chapter will be as follows. After a brief statement of the general case for externalism, I distinguish three basic varieties of so-called ‘causal’ theories in the literature on the subject: actualist, covariational, and teleological theories. A discussion of certain difficulties that have been noted in connection with these theories, with particular attention to the problem of misrepresentation, will lead me to explore the possibilities of a teleological account of mental content. I end this chapter with raising some doubts about the prevalent forms of causal theory. Is externalism really able to escape from the philosophical obsession with intrinsic content?

1. A general argument for externalism

The relational nature of cognitive states

The single most important general argument in support of externalism, as I understand it, is based on the fact that cognitive phenomena are essentially relational in nature. Cognitive states determine a relation between subject and object of cognition; while cognitive states themselves are generally considered to ‘belong to’ the subject, the objects to which they relate are typically states of affairs in the subject’s environment. If cognition is essentially relational in this sense, the individuation of explanatory kinds in cognitive science must advert to factors that are external to the individual subject; in particular, external truth conditions will play an important role in determining the nature of the subject’s cognitive states. Therefore, some form of externalism must be correct.
It is this relational aspect of cognition that played a pivotal role in several of our earlier discussions as well. Thus, in chapter three I argued against eliminativism that folk psychology is not so much a theory of what goes on inside the subject, but rather a way of descriptively relating the subject to his environment. On the dual aspect theory proposed there, the propositional attitudes define real states of the subject that are logically captured as semantic relations between subject and world. Similarly, in chapters four and following I drew attention to the fact that a properly epistemic understanding of neural nets requires that we attend not so much to their intrinsic properties, but rather to the specific contribution these properties make to enabling the comprehensive system to organize its interaction with the environment, or, in other words, in this sense to use the representations instantiated by states of the brain. Finally, our discussion of internalism in chapter six revealed that a ‘narrow’ construction of mental states tends to lead to the elimination of all content, thus depriving the internalist approach of all cognitive significance (the Fodor-Stich paradox).

The type of relatedness envisaged by the above argument is quite different from the relational character of cognitive states acknowledged by internalist theories. Fodor, for example, is quite explicit in stating that propositional attitudes are relational in the following sense.

“On the one hand, mental states are distinguished by the content of the associated representations, so we can allow for the difference between thinking that Marvin is melancholy and thinking that Sam is (…) ; and, on the other hand, mental states are distinguished by the relation that the subject bears to the associated representation (so we can allow for the difference between thinking hoping, supposing, doubting, and pretending that Marvin is melancholy)” (Fodor 1980 (1981, 226), italics in original; for other statements of the same view see, for example, Fodor 1985a; Fodor 1990d, 312ff).

When this view of the propositional attitudes is combined with a computational account of cognition, the result is a theory on which mental states are computational relations between a subject and his symbols, the contents of which are defined in terms of computationally salient properties of the symbols. Allowing for a distinction between narrow content and broad content, Fodor’s position may be summarized as the claim that
“propositional attitudes are computational relations to symbols encoded in the brain, whose broad content is determined by the properties onto which they lock, and whose narrow content consists in a disposition to so lock. Psychology consists in stating laws about such dispositions” (Loewer and Rey 1991b, xxx).

The *internal* relations envisaged by Fodor are of an entirely different order from the *external* relations introduced above. For one thing, Fodor’s idea of relatedness adverts primarily to the *attitudinal* aspect of propositional attitudes, whereas the externalist account adverts primarily to their *content*. Moreover, Fodor expects his theory to be able to elucidate all relevant distinctions in content in terms of these internal computational relations alone. The present externalist account takes the exact opposite stance: its project is to understand internal computational structure from differences in content, defined in terms of the subject’s different relations to external states of affairs.

In taking the externalist view of cognition here, my framework is that of the dual aspect theory of the propositional attitudes. As was argued in chapter three, the cognitive states described by folk psychology are best seen as defining primarily a relation between subject and world, rather than one between a subject and his internal symbols. If you believe that it is raining, for example, this is not because a relation is established between the central, logical ‘I’ inside of you and a symbol that is tokened in your mental economy. Rather, a relation is established between you and a wet part of the world. Of course, cognitive science also makes the *additional* claim that this relation is realized by your tokening of a symbol that relates you to that part of the world. These two aspects should be sharply distinguished, however. The dual aspect theory allows us to distribute the aspects along different dimensions: the *logical* or epistemic aspect is the relation, while the *real* or natural part is the internal symbol instantiating the relation. Now, if our aim is to understand cognitive phenomena from a cognitive point of view, we are *equally* committed to both aspects. That is to say, we must study internal activity (which is real and, according to our best theories, computational in character) in terms of the subject’s external relations to reality (which are logical or epistemic). Hence, we must take an externalist view of mental representation.

*Do we still need mental representations?*

It is apposite at this point to insert a note on the status of mental representations under externalism. The externalist notion of mental content is clouded by
a possible misunderstanding that is better nipped in the bud. Some readers probably felt slightly uneasy with the above remarks. It may have occurred to them that the very notion of mental representation is threatened with redundancy once the focus of interest is shifted from internal processing to external relations. In the context of internalist theories of mind, mental symbols served a clear purpose. They were needed to go proxy for states of affairs in the external world: the world as such may not be accessible to the subject, but its mental symbols are immediately present to consciousness, affording to the subject mediate (inferential) access to the world. Under the operation of externalism, by contrast, the primary relation is not between subject and symbol, but between subject and world. Hence, it would seem that the world itself is immediately ‘given’ to the subject. Therefore, mental representations are no longer needed: they do not add to our understanding of cognition. Externalism thus seems to defeat the purposes of any cognitive science whose central explanatory construct is the notion of mental representation.

Depending on one’s philosophical credentials, the above conclusion may be considered as either a virtue or a vice of externalism. On the one hand, internalists will be inclined to argue that the externalist approach badly trivializes the explanatory aims of cognitive science. An argument along these lines is developed by Fodor and Pylyshyn in their well-known attack on Gibson’s theory of ‘direct perception’ (Fodor and Pylyshyn 1981; Gibson 1979; cf. Koenderink 1980). I shall come back to the controversy over Gibson’s theory in the final chapter. Die-hard externalists, on the other hand, will be prepared to bite the bullet; they may point out that their approach constitutes indeed a radical break with the misguided philosophy of Cartesianism and with the ancient ‘Way of Ideas’. An example of this attitude is probably Bertrand Russell’s, who in his doctrine of ‘direct acquaintance’ energetically opposed the assumption that “if anything is immediately present to me, that thing must be part of my mind” (Russell 1956, 147). “On the contrary,” Russell professed, “I hold that acquaintance is wholly a relation, not demanding any such constituent of the mind as is supposed by advocates of ‘ideas’” (Russell 1912, 161; see also Peacocke 1983, ch. 7).

Although the externalist response strikes me as the more plausible of the above reactions, I think that neither of them are really appropriate. The lesson to be drawn from the relational nature of cognitive states should emphatically not be that mental representations are redundant. Rather, the conclusion should be that ‘having knowledge’ is primarily a matter of relating to the world, not of being related to symbols. As we have seen, traditional epistemol-
ogy has always evinced a tendency to analyze knowledge, as a relation between subject and world, in terms of a new and quasi-cognitive relation between subject and symbol. The upshot of the above argument is that this tendency should be resisted. Mental symbols are just \textit{not known}; they are in no way themselves the objects of knowledge, let alone that they should be the primary objects or \textit{prima nota}. To construe mental content as that which is immediately ‘given’ to the subject in internal consciousness, in a sense in which the external world is \textit{not} ‘given’, would in fact be a mere reduplication of the relation that constitutes knowledge.

Still, knowledge is realized \textit{by means of} internal representations: for a subject to be cognitively related to the world is for him to be tokening mental representations that go proxy for external states. The dual aspect theory of the propositional attitudes may once more serve to put this point into perspective. If individuals bear a certain relationship to reality that we logically identify as \textit{knowledge}, this does not make their relationship to reality any less \textit{real}. In particular, as was argued in chapter three, the nature of propositional attitudes, as describing a relationship between subject and world, is perfectly compatible with the claim that this relationship is realized in part by means of intricate computational mechanisms which can be studied from a naturalistic point of view. It would not be incorrect to say that mental representations are what makes knowledge \textit{psychologically real}. The task of cognitive science is to study this psychological reality from an epistemic point of view, that is, by determining for any given computational process not just how it is internally structured, but how it contributes to the subject’s epistemic relation to reality. When viewed in this light, externalism appears as a perfectly natural extension of the dual aspect theory: by letting truth conditions enter into the individuation of mental representations, it elegantly manages to combine the aspects of logic and reality.

\textit{Some ramifications of the general argument}

The relational argument presents the case for externalism in quite general terms. It is best seen as a generic argument that hosts a family of more specific arguments, each highlighting a different aspect of externalism. Some important ramifications of this idea were discussed in the previous chapter. Let me briefly rehearse two or three characteristic examples here.

In the first place, we found the argument from supervenience to be seriously flawed. It failed to establish internalism because it overlooked the possibility of a supervenience base that extends beyond the subject’s local physical
structure. Although intended officially as an argument from physicalism, the appeal to supervenience turned out to depend on a Cartesian premise that begs the question of internalism, namely, that mental content is determined exclusively by features intrinsic to the subject. Furthermore, so I argued, due consideration of the relational nature of cognitive states throws serious doubt on the internalist interpretation of supervenience. If knowledge is relational in the sense outlined above, the principle of supervenience is more readily understood as an argument for externalism, according to which the subject’s epistemic relation to reality supervenes on causal relations between organism and physical environment.

A similar strategy was followed for defusing the argument from functionalism. Granted that mental states are functional states, this does not necessarily mean that their individuation advertes only to factors internal to the individual. Functionalism evinced a fundamental ambiguity at this point, as functional states may be either narrow or wide. The case for internalism was based on the assumption that functional states are necessarily narrow. If externalism is correct, however, the web of functional relations extends beyond the virtual skin of the mind. I argued that the relational nature of cognitive states is more readily combined with wide functionalism than with a narrow construal of functional states.

Finally, both the arguments from epistemology and from computationalism pressed the idea that the subject of cognition is, in a way, locked ‘inside’ the organism, hence ‘outside’ the world. The primary relation is to internally defined symbols, from whose ‘narrow content’ the subject puts together his view of the external world. According to this line of thought, to understand knowledge is to take the internal subject’s point of view. It follows, then, that the computational profile of cognitive states should be determined independent of any external factors (argument from computationalism), and that, on pain of circularity, no specific knowledge about reality should be presupposed in the explanation of cognition (argument from epistemology). The fundamental difficulty with this defense of internalism is that it ignores the relational nature of cognitive states. I think it is simply a mistake to think of internal symbols as being ‘given’ to the subject, or to assume that the subject’s primary cognitive relation is to something like ‘intrinsic contents’. If we acknowledge the fact that cognition is essentially a relation between subject and world, we may free ourselves from the philosophical obsession with the Cartesian subject. The so-called ‘problem of the external world’ and the purported circulus in probandi, for example, evaporate once we stop thinking of ourselves as subjects tucked
away behind the veil of our senses. If there is a ‘problem of knowledge’ at all, it is how organisms use internal representations to reliably organize their interaction with reality. And this problem surely demands that we take some form of externalist position with regard to mental content.

These remarks may suffice as a general introduction to externalism. In the next section I present the three basic types of externalist theory of mental content, while the remainder of this chapter will be devoted to a discussion of certain difficulties that have been noted in connection with externalist theories.

2. Causal theories of representation

Three types of causal theory

Persuaded by something like the above line of reasoning, many philosophers of cognitive science have turned to an analysis of the relations the mind/brain bears to its environment. Within a broadly naturalistic framework, this approach invites a causal analysis of mental representation: its purpose is to study the way in which logico-epistemic relations between mind and reality are instantiated as real relations consisting of various causal components. Figure 7.1 illustrates the project in rough outline. Minds are linked to reality by causal chains that may be either mediated or unmediated. Intuitively speaking, the idea is that unmediated contact takes place when a subject $S_1$ is causally affected by the distal object itself, while contact is mediated when a subject $S_2$ is affected by a mere representation of the actual object. When you are looking at an apple, for example, your sensory surface is affected by the ambient array of energy issuing directly from the object itself through a causal channel $C_1$. Hitting your sensory surface, the incoming flow of energy is transformed and submitted to computational processing, resulting in the tokening of a neural representation $R$ with the content ‘apple’ (figure 7.1, top). Now, on the externalist approach considered here, the task of cognitive science is to establish how the various components of this causal chain (input channel, transduction, internal processing) contribute to the subject’s overall epistemic relation to reality, and to the determination of $R$’s representational content.

If, subsequently, you communicate the content ‘apple’ to others by telling them you saw an apple, or by showing them a picture of it, your representation of the apple will cause them to token the same mental content (figure 7.1, bottom). Intuitively speaking, the factors determining this latter content go back on your original contact with the apple. In this case, then, the causal chain between the apple and its mental representation by others is mediated, involv-
ing two distinct causal channels $C_1$ and $C_2$. I will have more to say on the
distinction between mediated and unmediated causal chains later in this chap-
ter. Suffice it here to point out that the central concern of externalist accounts
is to analyze the causal relations between the mind/brain and its distal objects.

At least three basic types of causal theory have been proposed in the recent
literature on the subject, namely, those that appeal to actual historical causal
chains, those that appeal to various sorts of causal covariation analyzed in
terms of counterfactuals, and those that appeal to some kind of ‘natural teleol-
ogy’ (for an overview, see Loewer and Rey 1991b, xxiii ff; Sterelny 1990, 114 ff;
Von Eckardt 1993, 214 ff; Cummins 1989, chs. 4-7). I shall briefly examine each
of them in turn.

**Figure 7.1: Causal theories of representation**

A rough outline of the externalist approach to content determination. The logico-epistemic link
between mind and reality is analyzed in terms of the causal chains instantiating this relation.
The relevant causal chains may be either direct (unmediated) or mediate. If they are direct, the
subject comes to have a token mental representation $R$ with content ‘apple’ by virtue of his
causal contact with the distal object itself (top). The chain is mediate if the subject comes to
token a mental content by virtue of an intermediate representation of the object (bottom). The
causal approach to mental content is obviously thoroughly empiricist in character. It presum-
ably entails that all determination of content is ultimately based on some unmediated contact
with reality. Notice that, for all its intuitive appeal, the distinction between mediated and
unmediated causal chains starts to break down if also the intermediate stages in allegedly
‘unmediated’ relations, such as retinal images, are analyzed as proper representations. I come
back to this problem later in this chapter.
1. Actualist theories

The first type of causal theory, which we may call the ‘historical’ or ‘actualist’ variety for reasons that will emerge shortly, was developed in the late 1960s and early 1970s, when several prominent philosophers of language rallied against the then prevalent doctrine of internalist semantics. Instead of the received view that the denotation of referring expressions in natural language is determined by their associated internal descriptions, they submitted that, for some expressions at least, reference is determined by causal links to the objects referred to (see, for example, Kripke 1972; Putnam 1975b). A distinctive feature of these ‘causal theories of reference’ is the fact that they envisage the actual causal chain linking a speaker’s use of an expression to an original user’s act of dubbing the object in question—which is why they may be called ‘historical’ or ‘actualist’ theories. Actualist theories were developed in particular for proper names and natural kind terms. Thus, Saul Kripke keenly argued that proper names and natural kind terms function as ‘rigid designators’: they designate the same object or substance in every possible world in which they designate anything at all. This steady reference is supposed to be fixed by the causal chain linking the speaker’s use of the expression to some object or substance in the actual world. For example, in the actual world the substance H\textsubscript{2}O was originally dubbed ‘water’. According to Kripke, the chain of speakers that extends from the dubbing event in the past to our present use of the expression ensures that by ‘water’ we mean H\textsubscript{2}O, while our Twins on Twin-Earth by the same expression will mean a different substance XYZ.

Actualist accounts of meaning are now widely recognized to be inadequate as a theory of mental content (cf. Loewer and Rey 1991b, xxiv). In the first place, they crucially involve events that require intentional characterization themselves, such as ‘dubbings’, ‘communications’, and ‘understandings’. For example, in order to establish a link between H\textsubscript{2}O and the expression ‘water’, the original dubber must have meant the water and not the fish swimming in it. Moreover, he must have been able to communicate this specific content, and his audience must have been able to understand it. When it comes to explaining these specific abilities, the question of mental content is simply begged.

Even waiving the above difficulty, actualist accounts appear to be unacceptably narrow in scope. For one thing, they are restricted to mental contents that are explicitly associated with linguistic expressions; it is not easy to see how the approach can be extended to other domains of cognition. Moreover, actualist theories may be readily applicable to uniquely referring expressions and certain natural kind terms, but it is difficult to see how they can be turned
into an account for predicates generally. Suppose, for example, that you want
to name a certain shade of blue. You look around for a suitably colored object,
point it out to your audience, and solemnly declare its name to be ‘X’. Now,
unfortunately, this hypothetical ceremony does little to help us understand the
factors that determine the mental content tokened by you and your audience.
The ceremony as such contains nothing to determine whether the content in
question is indigotic, blue, or color, nor does it rule out the possibility that the
intended content is the colored object, or perhaps even the fact that the color
strikes you as agreeable or hideous. The actual path from ostension to representa-
tion simply fails to discriminate between these possible contents. The diffi-
culty envisaged here may be called the qua-problem (cf. Sterelny 1990, 116ff):
concentrating on the single actual chain from object to mind, actualist theories
are necessarily indeterminate with regard to the various possible descriptions
or aspects under which the ostended object or state of affairs may be repre-

The qua-problem suggests that something more than actual dubbing routes
is needed. In particular, it suggests that we should attend to the set of counter-
factual routes ruling any given representation—a subject’s liability to use the
same expression (hence to token the same content) under different circum-
stances. Much the same conclusion may be gleaned from a consideration of the
possible world semantics involved in Kripkean theories. Recall Kripke’s claim
that some expressions function as rigid designators: they pick out the same
object or substance in all possible worlds in which they pick out anything at
all. Now, we may ask, what is the ‘semantic glue’ that makes these expressions
stick to substances or objects in this way? The fact that there is an original
causal route in the actual world does not by itself explain the expression’s li-
ability to attach to the same or similar substances in other possible worlds.
Similarly, the dubbing ceremony as such does not explain our cognitive capac-
ity to extend the use of the expression beyond its original application in the
dubbing ceremony. If we try to spell out the semantics behind Kripke’s claim,
we find that it is not so much the actual causal chain from ostension to repre-
sentation that matters, but rather the set of laws or reliable correlations be-
tween possible circumstances and correct uses of the expression.

Actualist theories are now virtually defunct. (For a recent effort to defend
a Kripkean type of causal theory, however, see Sterelny 1990, ch. 6; cf. Von
Eckardt 1993, 215ff.) In the remainder of this chapter I shall concentrate on
covariational and teleological theories of mental content.
2. Covariational theories

The difficulties with actualist theories led philosophers in the late 1970s and early 1980s to explore the idea that the relation between symbol and object is counterfactual rather than actual (see, for example, Stampe 1979; Dretske 1981; Stalnaker 1984; Fodor 1984, 1986). The resulting new type of causal theory has variously been called the ‘indicator’ approach (Sterelny 1990, 118ff; Von Eckardt 1993, 218ff), the ‘covariational’ approach (Cummins 1989; Loewer and Rey 1991b, xxv), or the ‘counterfactual’ approach (Fodor 1987; Loewer and Rey, loc. cit.). The distinctive feature of this type of theory is the claim that content is determined not by the actual causal ancestry of mental states, but by the nomological ‘slot’ filled by these states in the maze of causal relations between brain and world. The covariational view reminds of the doctrine of wide functionalism, of which it is indeed a specific extension. As we have seen in the previous chapter, wide functionalism entails that mental states are type-identified in terms of their ‘wide’ i/o control, that is, in terms of their role in causally mediating between distal factors, other internal states, and behavioral effects. The covariational approach proposes, in effect, to obtain an account of mental content from the nomological character of these ‘wide’ relations.

To a first approximation, a rough version of the view may be stated as follows:

“the symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings” (Fodor 1987, 99).

The necessary and sufficient conditions for this ‘reliable causation’ are expected to be specifiable in terms of lawlike correlations between instances of properties of distal objects and tokenings of mental representations. A first effort in this direction would be to endorse the claim that

“a symbol expresses a property if it’s nomologically necessary that all and only instances of the property cause tokenings of the symbol” (Fodor, op. cit., 100).

This hypothesis, which Fodor has called the ‘Crude Causal Theory’, is obviously too strong. Problems arise both with regard to the all part of the conditional and with regard to the only part. In the first place, it is simply false that all instances of properties cause mental tokens of the appropriate type. Having
the property of ‘being a table’, for example, is not a sufficient condition for an object to cause me to token the content ‘table’. In point of fact, most tables do not cause me to have ‘table’ contents, simply because most tables are not present to me. Even the table I am writing at now often fails to evoke the appropriate content; under certain conditions I am liable to mistake it for something else, and most of the time I am simply not paying attention. To be sure, the problem may be solved by explicating the reference to ‘nomological necessity’ in the original claim, adding a counterfactual clause to the Crude Causal Theory: all P’s would cause tokenings of ‘P’ contents if only certain conditions C₁, C₂, ..., Cᵢ are met.

As for the second part of the Crude Causal Theory, it is also false that only instances of a property P may cause tokenings of their associated content. Sometimes we make mistakes, misrepresenting a property that is not-P as being P. To account for this possibility, the intuitive response is again to take recourse to a counterfactual clause, stating that only P’s would cause tokenings of ‘P’ if certain conditions D₁, D₂, ..., Dᵢ are met.

3. Teleological theories

If the problems with covariational theories are handled along the lines indicated here, a certain amount of idealization is introduced.² The emerging theory claims in effect that, under ‘ideal’ (‘optimal’, ‘normal’, etc.) circumstances, the appropriate contents will be reliably caused by their distal objects. The fundamental problem is then how to define these circumstances in a principled and non-question-begging way. What distinctive features are shared by all and only the different causal chains that may lead from instances of a distal property P to the tokening of representation R, such that R may be said to ‘reliably co-vary’ with P? To be sure, both property P and representation R are shared, but this answer is obviously not very informative. In recent years, the problem of idealization has fostered a new species of causal theory, which has become known as the ‘teleological’ approach (see, for example, Fodor 1984 and 1990d; Millikan 1984, 1989, 1990, and 1993; Papineau 1987 and 1993, ch. 3). Teleological theories try to explain the notion of ‘ideal’ or ‘normal’ circumstances in terms of the ‘natural function’ subserved by mental representations. The basic idea underlying the approach should be familiar from previous chapters: arguably, the content of internal representations is constrained by the organism’s use of these representations in organizing its interaction with the environment. Now, teleological theories propose to analyze the notion of mental content specifically from the point of view of this purposiveness. Whether
or not a given representation is useful for a given organism, and if so, in what sense, depends on the teleological structure of its interaction with reality. This structure comprises several layers of potentially relevant factors, including the specific needs posed by the organism’s body, the organization of behavior at a social level, and the requirements defined by ecological characteristics of the natural habitat.

Methodological provisos

In the next section I turn to a more ample discussion of the covariational and the teleological approach. As my angle of approach I choose the problem of misrepresentation in causal theories. I do so for two reasons: first, because this problem is without doubt the best studied objection to causal theories, which has drawn much attention in recent literature, and secondly, because I think it divulges much of the internal dynamics of causal theories, helping us to understand in particular the rationale behind the development of teleological theories of mental content.

Most of my examples in what follows will be recruited from the domain of perceptual beliefs. This restriction calls for a word of justification. As Barbara Von Eckardt has pointed out, proponents of causal theories generally tend to concentrate on a ‘target set’ of representational states, states to which the basic ideas of causal theories seem to apply in a relatively straightforward manner (Von Eckardt 1993, 218). Thus, most causal theories attend primarily to de re representations of simple subject-predicate form, such that objects O are represented as having property P. In addition, causal theories are primarily concerned with the fixation of perceptual beliefs, as opposed to non-perceptual beliefs and desires. I will adopt these customary restrictions in what follows. My reasons for doing so are twofold. In the first place, if it turns out that causal theories are unable to deal with even the simpler kinds of mental representation, this sufficiently demonstrates their inadequacy. Secondly, however, if it can be shown that some form of causal theory is acceptable as an account of perceptually caused beliefs, there is reason to believe that it may eventually also be able to deal with more problematic kinds of mental states.

3. The taming of the shrew

Misrepresentation and the disjunction problem

The problem of misrepresentation in causal theories of content arises from the fact that the conditions imposed on representation comprise the representation’s
in principle, therefore, nothing can be a representation unless it is also veridical. For example, nothing counts as a representation of a mouse unless it is caused by mice; but then it is ipso facto a veridical representation of a mouse qua mouse. Apparently, there can be no such thing as misrepresenting mice as something else, or something else as a mouse.5

For convenience of exposition, let me adopt the notational device introduced by Cummins (1989, 56) for naming mental contents. By definition,

1. $|A|$ =def type of mental representation denoting objects as having property A.

Equivalently, I shall use ‘$|A|$’ to name the type of mental content representing the putative fact that a certain object has property A. For example, tokenings of $|\text{blue}|$ represent objects as being blue. Similarly, $|\text{mouse}|$ tokens represent the putative fact that mice are present. Presuming that neural/mental representations are type-individuated in terms of their content, I shall also use ‘$|A|$’ to refer to those neural structures R that mean $|A|$. The problem of misrepresentation can now be stated as follows. According to causal theories of content, tokenings of $|A|$ are caused by instances of A. The possibility of error requires, however, that $|A|$ tokens are sometimes caused by B’s that are not-A. The possibility of error is very real; therefore, causal theories must be wrong.

The intuitive response to this objection is to point out that misrepresentations of B as $|A|$ may indeed be caused by B, but that this causal link is certainly not reliable. A-caused $|A|$ tokens, by contrast, are reliably caused by instances of A. There is a vast difference, so one might want to argue, between $|A|$ tokens that are ‘robust’ and those that are ‘wild’. Robust tokens are reliably caused by whatever they represent, whereas wild tokens misrepresent their causes.

Stated in this blunt form, the intuitive solution is inadequate. First of all, recall that our aim here is to explain why the A-to-|A| link is reliable and the B-to-|A| link is not. The present reply merely restates the question in new terms; if we ask what makes some tokens robust and others wild, the question of reliability will be begged. Moreover, and more importantly, we should keep in mind that causal theories are intended as an account of the determinants of mental content. They seek to explain the fact that a mental representation R has content $|A|$ in terms of R’s being caused by A. Now, if some R tokens are caused by A’s while others are caused by B’s, R tokens may be caused by either A’s or B’s. Apparently, then, R represents the disjunctive content $|A$ or $B|$. Again, we find that misrepresentation is ruled out.
The latter objection has come to be known as the *disjunction problem*. Its consequences for the notion of mental content in causal theories are truly devastating. Our capacity for error is such that, depending on the circumstances, almost any property may cause us to token almost any representation. Unless causal theories find a way to control this gross semantic indecision, they will end up conflating all mental representations onto a single, disjunctive content of indefinite length, \( A \) or \( B \) or \( C \) or \( \ldots \).

The disjunction problem calls for a principled and independent way of distinguishing between the robust and the wild. As we have seen in the previous section, the robustness of representations is arguably a function of their nomological profile. Robust \( \text{mouse} \) tokens covary with mice in a fashion that is not shared by their wild counterparts. The question, then, is how to spell out this distinctive aspect of ‘veridical covariation’. In the next couple of paragraphs I examine three important attempts to answer this question.

**Of \( \text{mouse} \)s and men**

The first sustained attempt to develop a covariational theory is Fred Dretske’s ground-breaking work on *Knowledge and the flow of information* (1981). Basing himself on the mathematical theory of communication developed by Claude Shannon and Warren Weaver in the late 1940s, Dretske sketches the outlines of

---

![Figure 7.2: Misrepresentation and the disjunction problem](image_url)

A simple case of perceptual misrepresentation occurs when shrews intrude on the mouse-to-R connection. Intuitively speaking, the content of the subject’s representation R must be \( \text{mouse} \). Accordingly, shrew-caused tokens of R are wild (misrepresentations), while mouse-caused tokens of R are robust (correct representations). Causal theories of content have trouble distinguishing between the robust and the wild: if content is determined by distal cause, then it is equally correct to say that R’s content is \( \text{shrew} \) as that it is \( \text{mouse} \) (because shrews cause \( \text{mouse} \)s, too), or that it is \( \text{mouse or shrew} \) (because both mice and shrews may cause \( \text{mouse} \)s). (Freely rendered after Cummins 1989.)
a fully naturalistic account of the causal conditions for belief fixation, with particular regard to perceptual beliefs. His analysis is framed in terms of the familiar sender-receiver model of perception. According to this model, signals containing information about their distal source are passed through a ‘communication channel’ between object and subject, eventually causing the subject to go into the appropriate representational state. The content of this state is determined by the informational content of the signal, which in its turn depends on properties of the source.

Figure 7.2 pictures a simple perceptual situation in which a subject comes to have a token representation $R$ with content $\,|\text{mouse}|\,$ in the presence of mice (top). Sometimes, however, it may happen that the observer tokens $\,|\text{mouse}|\,$ in the presence of shrews (bottom). The latter is a case of misrepresentation. Dretske’s account of what happens is roughly as follows. Mice and shrews signal their presence to all observers, probably by means of an ambient array of light travelling from object to subject. Let us call this signal S. Suppose, for the sake of argument, that the subject can see only the animal’s tail. Speaking very loosely, we may then say that it is the tail which carries the information that a mouse is present; in this sense, the tail is the signal S. Now, S causes the subject to token $R$. What is more, $R$’s content is determined by the information contained in S. The disjunction problem looms large again: while $R$’s content is inherited from S, S itself inherits its content from properties of the distal source, that is, from mice or shrews.

Dretske’s work contains at least two distinct leads for grappling with the problem of misrepresentation, only one of which is pursued by Dretske himself, however. The first lead may be taken from the notion of ‘information contained in the signal’, of which Dretske submits the following definition:

“Informational content: A signal $r$ carries the information that $s$ is $F = \text{the conditional probability of } s’s \text{ being } F, \text{ given } r \text{ (and } k\text{), is 1 (but, given } k \text{ alone, less than 1)}$” (op. cit., 65).

The variable $k$ in this definition stands for what the receiver already knows about the possibilities existing at the source; I shall ignore this magnitude for present purposes. Applying the above definition to the mouse/shrew example, we find that the signal (tail) does not carry the information that mice are present, for the tail-mouse correlation is evidently less than perfect. By contrast, the correlation between tails and mice or shrews is 1. Hence, on this account, the informational content of signal and representation is disjunctive almost by definition.
When viewed from this perspective, the problem of misrepresentation appears to be caused by the requirement that the relevant conditional probabilities must be perfect. Perhaps this requirement can be toned down, however. Perhaps content is determined by correlations that are less than perfect, or that are perfect only if certain additional constraints are taken into account, such as a restriction with regard to environmental context. Here lies a possible opening toward solving the disjunction problem. Thus, it is quite conceivable that the conditional probability of something’s being a mouse, given the tail, is significantly larger than the conditional probability of its being a shrew. Put very crudely, it may well be that normally every tail is a mouse tail—a fact that may then be used to motivate the distinction between wild and robust |mouse| tokens. Whether this regularity actually holds in any given type of context is an empirical question, of course, to be settled only by studying the actual conditions under which specific subjects make the acquaintance of tails.

As noted in the previous section, the reference to environmental constraints such as normal conditions introduces a teleological element into the causal approach. Dretske himself does not consider this possibility in his earlier work. In later publications, however, he has considerably modified his original theory in a teleological sense, with particular regard to the problem of misrepresentation (see, for example, Dretske 1986 and 1988, 64ff). I return to the teleological amendments later in this section.

The second lead for solving the mysteries of misrepresentation, and the one explored by Dretske himself, lies in the field of concept formation (Dretske 1981, 190ff). Arguably, the content of many of our representations was acquired through a process of conditioning by our epistemic peers. Now, the conditions imposed on the acquisition of a representation are also the conditions that determine its content: which concept a subject has learned depends crucially on the circumstances under which he was trained to token it. For example, if you are trained to token a representation R in the presence of certain instances of green, yellow, and red fruits which are all apples, and if you are trained further to discriminate these from oranges, pears and other non-apples, the acquired set of discriminatory and identificatory dispositions comprised by R may properly be said to bear the content |apple|.

Dretske’s explanation of misrepresentation is based on the idea that mental content is fixed during the learning period. Pupils are exposed to perceptually evident cases of A selected by teachers, who encourage them to token |A|’s. If, when the learning period is over, |A|’s are occasionally tokened in the presence of non-A’s, these tokens are misrepresentations by virtue of the correlation established in the learning period.
Probably the most fundamental difficulty with Dretske’s proposal is that it ignores the relevant counterfactuals for determining which symbol-to-world correlation the training has brought about (cf. Fodor 1987, 104; Cummins 1989, 67-69; Sterelny 1990, 122). As we have seen, causal theories generally entail that the connection between $|A|$ and A’s is one of nomological dependency. The connection depends not only on what actually happened during the learning period (exposure to A’s), but also on what might have happened. Now, by hypothesis, some B’s that are not-A may occasionally cause tokens of $|A|$. Dretske assumes that this did not happen during the learning period. Yet, it might have happened; the acquired concept would then be $|A|$ or $|B|$ rather than $|A|$. Therefore, nomologically speaking, that is with regard to the relevant counterfactuals, the acquired concept will be the disjunctive content $|A$ or $B|$.5

A faster and more intuitive argument may drive this point home. Teachers are only human; they are liable to make mistakes. They may occasionally token $|A|$s in the presence of B’s. If this happens when they are teaching, pupils will learn to respond to A’s or B’s. Hence, on Dretske’s own principles, pupils are bound to develop a disjunctive concept $|A$ or $B|$, which is destined to grow like a semantic snowball. As pupils become teachers, the disjunction is passed on to future generations. Disjunction will spread like a cognitive virus.

Although the problem noted here is very real, Dretske’s idea should not be dismissed too lightly. As Kim Sterelny has pointed out (1990, 122), the operation of control and feedback mechanisms (some of which are teachers) is decidedly relevant for the fixation of mental content. They serve to tune the organism’s response to environmental factors. If we are looking for an explanation of why the A-to-$|A|$ correlation is reliable, control mechanisms may be at least part of the answer. These mechanisms are not infallible, to be sure. The correlations they establish typically fail to meet Dretske’s requirement of perfect conditional probability, discussed above. As it stands, however, this requirement is probably too strong. Regularities can be reliable in spite of the fact that they break down in some logically or physically possible world. A defense of Dretske’s position along these lines suggests an attenuated notion of reliable covariation, such that representations are reliable if they are tuned to actual or normal conditions in the organism’s environment.

Horses, wild $|\text{horse}|s$, and cows

An alternative solution for the problem of misrepresentation is proposed by Fodor (1987). His account draws on the observation that our ability to make
mistakes is parasitic on our ability to get things right. In more solemn words, “falsehoods are ontologically dependent on truths in a way that truths are not ontologically dependent on falsehoods” (Fodor 1987, 107). We saw that a solution of the disjunction problem calls for a principled distinction between wild tokens and robust ones. Fodor takes up the challenge in the following, slightly confusing, argument.

"From a semantic point of view, mistakes have to be accidents: if cows aren’t in the extension of ‘horse’, then cows being called horses can’t be required for ‘horse’ to mean what it does. By contrast, however, if ‘horse’ didn’t mean what it does, being mistaken for a horse wouldn’t ever get a cow called ‘horse’. Put the two together and we have it that the possibility of saying ‘That’s a horse’ falsely presupposes the existence of a semantic setup for saying it truly, but not vice versa. Put it in terms of CCT (the Crude Causal Theory, JS), and we have it that the fact that cows cause one to say ‘horse’ depends on the fact that horses do; but the fact that horses cause one to say ‘horse’ does not depend on the fact that cows do” (op. cit., 108).

The upshot of this argument is that the causal connection between cows and |horse|s is, as Fodor calls it, asymmetrically dependent upon the causal connection between horses and |horse|s. No cow-to-|horse| connection without a horse-to-|horse| connection; but the horse-to-|horse| connection does not presuppose the cow-to-|horse| connection. Apparently, then, we have found a way to distinguish robust tokens from wild ones. B-caused |A| tokens are wild only if they are asymmetrically dependent upon A-caused |A| tokenings. Conversely, robust tokens are asymmetrically independent from wild tokens. And to round things off, disjunctive tokens are symmetrically dependent upon one another: A-caused |A| or |B|s require B-caused |A| or |B|s, and vice versa.6

There is something strange about this formal solution to the disjunction problem. Exactly what is wrong with it is best illustrated by considering a counterexample discussed by Fodor himself. Suppose a subject S learns the concept |horse| by being trained on a set consisting of exclusively wrong specimens, for example horse-like cows. If S subsequently applies this concept to horses, his use will be correct, for the acquired concept is by hypothesis |horse|. Yet, the fact that S now correctly applies |horse| to horses is a result of his previous misrepresentation of cows as |horse|. “Had he not previously applied ‘horse’ to cows, he would not now apply ‘horse’ to horses. So it looks
like we’ve got error without asymmetric dependence” (op. cit., 109). Fodor’s reply to this objection is to distinguish between two kinds of asymmetry, synchronic and diachronic. The asymmetric dependency required for wildness, Fodor claims, must be *synchronic*. In the example, however, the dependency of correct use upon wild use is *diachronic*: “my present disposition to apply ‘horse’ to horses does *not* depend on any corresponding *current* disposition to apply it to cows” (loc. cit.). Hence, the counterexample fails.

Fodor’s reply strikes me as less than satisfactory. I want to mention three difficulties here. In the first place, the sudden restriction that asymmetry must be synchronic is suspiciously *ad hoc*. No independent motivation of this move has been given, and there is reason to expect that none can be given in principle. Moreover, and more importantly, Fodor’s reply entails a strange imbalance between a subject’s past and present dispositions. It suggests that the subject’s *former* disposition to apply |horse| to cows has somehow ceased to exist (there is no corresponding *current* disposition), and that it has been replaced by a new and different disposition to apply |horse| to horses. But how can this be? Causality has not changed, nor have the subject’s dispositions. It is really the *same* disposition (the same mental representation) that was first triggered by cows and that is now triggered by horses. If cows were able to cause |horse|s in the past, they probably still can do so now. But this means that S’s current disposition to apply |horse| to horses depends (also) on his current disposition to apply them to cows, for the past disposition and the present disposition are really the same. Therefore, we have synchronic dependence of truth upon error after all, and Fodor’s solution is invalidated.

Finally, Fodor’s reply rests on an assumption that is vastly implausible. Fodor assumes that, in the case imagined, the acquired content will be |horse| in spite of the fact that the subject was trained exclusively on cows. But, surely, in a case like this it is far more plausible to suppose that the acquired concept R is |cow|, and that its later use for horses is erroneous. Alternatively, even the suggestion that R has the disjunctive content |cow or horse| is more plausible than Fodor’s original hypothesis. Fodor thinks he can skirt these alternative hypotheses by simply *stipulating* that the more basic relation is between R and horses. This apriorism obviously begs the question of content, however, which makes it unacceptable as a solution of misrepresentation.

Actually, this result was to be expected given the nature of Fodor’s approach. Dependencies between world-to-content relations can be stated only *after* the relevant contents have already been determined. Therefore, asymmetric dependency cannot serve to determine *what* the content of a (mis)repre-
sentation is. Conversely, if content is fixed a priori, the appeal to asymmetric dependency reveals itself as mere window dressing. The deeper reason behind the apriorist strain in Fodor’s theory is doubtless his general internalist outlook. As we have seen in the previous chapter, Fodor tends to think of mental content as being fixed even before external relations are established. This idea is captured by his notion of narrow content, defined as a disposition to lock onto specific properties in specific contexts. If external relations (properties and contexts) are bracketed, this internal disposition or ‘narrow content’ remains. Apparently, then, it is this residual ‘intrinsic’ content which determines that the basic relation of representation R in our example is with horses and not with cows.

Toward a teleological solution

If recourse to intrinsic content is denied, the solution for misrepresentation must be sought elsewhere. One particularly attractive avenue of approach is indicated by teleological theories. We have already seen several suggestions in that direction. Thus, a certain amount of teleology was entailed by idealization in counterfactual theories. Moreover, difficulties with Dretske’s stochastic analysis of informational content invited a consideration of the ecological constraints on certain conditional probabilities. Finally, we found that feedback and control mechanisms operating during learning are instrumental in tuning the organism’s response to environmental factors, which may partly account for the existence of reliable A-to-|A| correlations.

Viewed from the perspective of misrepresentation, teleology begins just where Fodor’s account left off: if Fodor singled out a ‘privileged mapping’ from narrow content to horse-contexts, teleological theories try to explain why this relation is privileged. Crudely but not inaccurately stated, privileged contexts are identified as the ones in which ‘normal’ conditions obtain, while privileged content is determined in terms of the representation’s matching ‘natural function’. These notions will become clear as we proceed.

Two caveats are in order here. In the first place, if I use Fodor’s notion of narrow content to state the purpose of teleological theory, this should not be misunderstood as saying that teleology presupposes narrow content. The explanatory framework of teleology is expressly externalist, or, more to the point, it is relational in the sense outlined in this chapter. It envisages both ends of the A-to-|A| relation, contexts as well as contents. Narrow content, by contrast, is supposed to be an abstraction over contexts: it is whatever ‘intrinsic content’ remains inside the subject if external relations are bracketed out. The
idea that such internal residue is in any sense contentful is explicitly rejected by the teleological approach. Secondly, if I say that normal conditions and natural functions lay down a privileged relation between mind and world, this should not necessarily be taken in the sense associated with traditional epistemology. Traditionally, epistemic privilege is explained by intrinsic characteristics of either mind or world. In the Cartesian tradition of internalism, the fact that representation R connects with property A is due to R’s having intrinsic content $|A|$. Alternatively, metaphysical realists may account for the same fact by claiming that the object is intrinsically of type A, and that it imprints itself on the subject under the $|A|$ form (as in the Aristotelian and Lockean metaphysics described by Putnam 1981, ch. 3). Teleological theory, as I understand it, rejects either of these alternatives: epistemic privilege is determined neither by context-as-such nor by content-as-such; rather, it depends equally on both sides of the knowledge relation. As we shall see, however, it is not always easy to strike a clear balance between the two sides.

According to teleological theories, the content of internal representations is constrained by its utility for organizing its interaction with the environment. Whether tokens are wild or robust depends on the circumstances under which they are tokened. To a first approximation, tokenings of representation R are robust if circumstances are normal, and they are wild if circumstances are abnormal. Correspondingly, R is said to represent whatever property would cause it under normal conditions, or, in slightly different words, normally-caused R tokens reliably indicate their distal causes. Misrepresentation occurs when conditions are not normal.

A simple analogy may illustrate this claim. Consider a measuring device such as a voltmeter, whose reading reliably indicates the electrical potential between the instrument’s terminals. If asked to explain this reliability, we readily point out that the meter is calibrated so as to correctly represent voltage if certain conditions $C_1$, $C_2$, ..., $C_i$ are satisfied. These conditions we call ‘normal’. They cover circumstances both inside and outside the instrument. Inside, for example, they demand that the central spring must be neither too tense nor too slack, that the circuitry is in working order, the terminals properly attached and free from corrosion, and so forth. Outside, the instrument must be free from undue influences, such as strong magnetic fields interfering with the meter’s normal operation. If any of these conditions is violated, the instrument’s reading will be a misrepresentation of the voltage between terminals, or at best a merely fortuitous representation. Under normal conditions, however, the meter yields a correct representation. By analogy, mental repre-
sentations are supposed to represent their causes under normal conditions. If normally a neural structure R is selectively sensitive to distal causes of type A, its content is properly identified as being \( |A| \). Under abnormal conditions, R may also be triggered by B’s that are not-A; R then misrepresents B as \( |A| \) by virtue of the normal A-to-R correlation.

There is a difficulty with this analogy, however. It may be objected that the voltmeter does not indicate voltage as such, but rather the ensemble of voltage and other relevant causal factors. That is to say, it represents not \( |\text{voltage}| \), but \( |\text{voltage or abnormal conditions}| \). If the meter’s dial indicates zero, for example, this means that either the voltage is zero or that conditions are abnormal (or both). The disjunction problem looms large again: a zero reading on a voltage of ten does not misrepresent the voltage, but correctly represents the disjunctive property ‘zero voltage or abnormal conditions’. By analogy, disjunction would also resurge with regard to neural structures considered as measuring devices.

In practice, the possibility of disjunctive content in measuring devices is not a problem at all, of course. When using a voltmeter, we know \textit{a priori} how to distinguish robust readings from wild ones. We know roughly which normalcy conditions must be satisfied for the meter to function as it should. These conditions follow from the instrument’s built-in teleology, from its intended function and its proper use, which are products of the instrument’s design. In the final analysis, the content of representations in artifacts such as voltmeters is determined by the original intentionality of their human designers. Philosophically speaking, the analysis begs the question of content, for the ‘original content’ of designers is left unexplained.

To escape from this quandary, the notion of teleology must be carried beyond the realm of human design. In the case of artifacts such as voltmeters, teleological structure is obviously \textit{derived}. If we are dealing with \textit{natural} cognitive systems, however, no such recourse is available: their teleology is ordained by nature rather than by man. What we need here is the notion of a \textit{natural teleology}, a teleology that does not derive from some other more basic form of intentionality, but that inheres in the texture of nature itself. Although this view has some difficulties of its own, it is at present the most promising approach available. It rests on the intuition that certain neural structures (representations) in organisms have a \textit{natural function}, namely, to enable the organism to organize its interaction with the environment, both physical and social. A representation’s utility, in this sense, depends on a number of factors, including the specific needs posed by the organism’s body, the organization of
behavior at a social level, and the requirements defined by ecological characteristics of the natural habitat. The natural teleological structure of cognition is defined by the ensemble of such factors.

In addition to being intuitively appealing, the notion of a natural teleology is supported by established ideas in modern biology, where functionality and purposiveness are basic explanatory means. Disciplines such as ethology and evolutionary biology, for example, study the way in which behavior and physiological structure are adapted to the needs and possibilities of organisms and their environment— an approach that is explicitly teleological. Results in these fields may lend support to kindred accounts of mental content. In the final analysis, we may expect the purposiveness of cognitive structure to be explainable in terms of adaptive success and survival value.

To give an example of the way in which natural teleology bears on content, I call once more on the case of the rattlesnake, introduced in chapter five. The nervous system of rattlesnakes contains a cellular structure R that is activated only if three distinct input conditions are met (see figure 7.3). First, R must get positive input from the snake’s infrared detectors, which under normal circumstances signal the presence of a nearby warm object. Moreover, the snake’s visual system must signal to R the presence of a relatively small object. Finally, the visual system must indicate that the object is moving. When R fires, certain hunting routines are activated and the rattlesnake will snap at the perceived object. Now, in addition to these physiological facts, we have ecological evidence about the rattlesnake’s natural habitat: normally, the simultaneous satisfaction of all three conditions means that small warm-blooded living animals are present; these animals are usually mice, on which the snake preys. In view of these facts, we may say that R is used by the snake as a reliable indicator of the fact that food is present: its natural function is to signal \textit{food}, or maybe even \textit{mouse}. Extending the analogy with measuring instruments, Paul and Patricia Churchland have suggested the term ‘calibrational content’ for this functional aspect of representation.

“The backbone of what we are calling \textit{calibrational content} is the observation that there are reliable, regular, standardized relations obtaining between specific neural responses on the one hand, and types of states in the world. The notion exploits the fact that specific neural responses are regularly caused by \textit{types} of state in the organism’s normal environment” (Churchland and Churchland 1983, 14).
To be sure, under abnormal circumstances the system’s reliability falters. This may happen under extreme temperatures, or when the snake is dislocated from its normal biotope. Experimental biologists (‘evil biologists’; cf. Putnam 1981, ch. 3) can easily fool the animal by moving hot rocks, or by manipulating infrared lamps. R’s natural function is not to be responsive to such abnormal conditions, however. Hence, non-food-caused R tokens are wild.

A specific cellular structure R in the nervous system of the rattlesnake is activated only if three distinct input conditions are satisfied simultaneously. The infrared system (A₁) must signal the presence of a nearby warm object, while the visual system must signal that the object is small (A₂) and that it is moving (A₃). When R is activated certain hunting routines are engaged: the snake snaps at the perceived object. Under normal circumstances in the snake’s natural habitat, the simultaneous satisfaction of all three input conditions is a reliable indicator of the fact that a small warm-blooded animal is near; these animals are usually mice, on which the snake preys. The natural function of representation R, in terms of its usefulness for the organism, is to contribute to the snake’s survival under normal conditions; this function is determined, on the one hand, by the snake’s bodily need for food, and, on the other hand, by the ecological profile of the snake’s natural habitat. In view of these teleological constraints, R’s content may be properly identified as \[\text{food}\], or maybe even more specifically as \[\text{mouse}\] (cf. Churchland and Churchland 1983; Newman and Hartline 1982).
For all its simplicity, the rattlesnake example illustrates the way in which the neural circuitry of cognitive systems in general can be understood as exploiting the regularities obtaining in their normal environment, each regularity corresponding to a specific neural representation or hypothesis about normal states of affairs. Complex systems such as the human mind presumably involve an astronomical number of such hypotheses. Moreover, they are presumably ordered in far more intricate ways than the hypotheses in our example, involving vast patterns of interrelated neural nets woven into a ‘society of mind’ (cf. Minsky 1986, esp. 309ff). Some of these structures will be innately specified like the fixed action pattern of the rattlesnake’s hunt, calibrated in a phylogenetic process of adaptive ‘learning’. Others will be the result of ontogenetic learning processes in individual organisms, where feedback and control mechanisms serve to tune the individual’s neural structure to the needs and possibilities of individual and environment.

Our brains are not equipped by nature to cope with all logically possible worlds, nor even with all physically possible worlds. The brain thrives on the fact that we live in what Margaret Boden (1984) has called a ‘cognitively friendly world’, a relatively small band of normal conditions within the broad spectre of possible worlds. The human visual system, for example, manages to extract useful information about the distal environment because it effectively simplifies the cognitive problem involved. Vision exploits

“certain computational constraints grounded in (cognitively friendly) aspects of the environment which might conceivably have been different. If the visual system assumes these aspects to be present, it will occasionally be subject to illusions but on the whole will interpret images reliably” (Boden 1984, 22).

In other conceivable worlds our cognitive system would simply be unable to serve its natural purpose. If all middle-sized physical ‘objects’ were engaged in continual and dramatic change, for example if chairs behaved like quantum particles, knowledge as we know it would not be possible. At the same time, however, it is important to bear in mind here the astounding plasticity of the human nervous system. Of all devices known to us it is by far the most versatile. The range of circumstances to which it can adapt is truly incredible. Still, it is finite. Notice that this approach to cognition is distinctly transcendental in character, an aspect to which I return in more detail in the final chapter.
Cognitive utility and functional misrepresentation

Our discussion so far has tended to concentrate on the environmental side of teleology, smart content being rooted in the friendliness of the world. As a result of this, it may seem that the only difference between teleological accounts of misrepresentation and earlier causal theories is the added normalcy constraint: robustness is determined not by correlation tout court, but by correlation under normal circumstances. Actually, however, teleological structure comprises much more than environment alone. Not normalcy but utility is the central notion. Whether a representation is useful for an organism depends on many other factors besides the friendliness of the world. It is a fully relational notion, involving conditions at the object side of knowledge as well as conditions at the subject side. To demonstrate this point, let me call attention to a well-known difficulty with teleological accounts, the problem of so-called ‘functional misrepresentation’ (see, for example, Taylor 1987; Sterelny 1990, 123ff).

If robustness is taken to depend on normalcy alone, misrepresentation under normal circumstances is impossible by definition. It has been noted, however, that the consequent of this hypothesis is false: misrepresentation under normal circumstances is a widespread phenomenon, which serves a perfectly natural function. Organisms may be calibrated to token $A$ even when, under normal conditions, this is most often a misrepresentation of some not-$A$ as $A$. Environments favoring this kind of misrepresentation are those in which it matters much less to get not-$A$’s right than not to get $A$’s wrong. Desert mice, for example, may be calibrated to token $\text{snake}$ at the merest ripple of sand: perhaps nine out of ten times the alert is false, but under circumstances it pays to be cautious. On the face of it, this seems to defeat the teleological account of misrepresentation, breaking the link between robust tokens and normalcy conditions, and/or that between normalcy and function.

To resolve the problem, two quite different lines of reasoning suggest themselves. The first is to try to somehow salvage the link between robustness and normalcy. Thus, one may be tempted to reconstrue the content tokened by the mice in the example so as to avoid the possibility of error under normal conditions. It may be argued that this content is not $\text{snake}$, for example, but rather $\text{potential danger}$. The underlying idea is that ripples of sand signal potential danger; therefore the mouse’s reaction under normal conditions is veridical without exception. I do not think this approach is very promising, however. In point of fact, I think it reinstates the disjunction problem in all its original vigor: if sand ripples are ‘potentially dangerous’, then why not also...
imitation rattlesnakes, or the shadows of clouds gliding over the sand, or other optical illusions that look like snakes? | potential danger | strikes me as a disjunctive content in disguise, which, when spelled out, is really tantamount to | snakes or sand ripples or clouds or illusions or … |. If this diagnosis is correct, the proposed solution inevitably leads to a view on which all misrepresentation is ruled out again, under normal conditions as well as under abnormal ones—a cure that is worse than the disease.

There is a more fundamental reply, however, which concentrates on the question of why sand ripples may be claimed to signal potential danger. More to the point, why is it at all useful for mice to token | snake |s, even when they are mistaken? It is certainly not useful to run away from sand ripples, nor to flee from the shadows of clouds. Rather, it is useful to avoid being eaten by snakes. The natural function of | snake | tokens is to enable mice to avoid being eaten. Therefore, ripple-caused | snake |s are wild even when they occur under normal conditions, for their natural function is to signal snakes, not ripples. If ripples in the sand cause mice to token | snake |s, this is only by virtue of their resemblance to snakes.

This second line of reasoning presses the insight that cognitive calibration takes more than normalcy alone. For a neural structure R to have content | A | it is not enough that it is caused by A’s under normal conditions. In addition, R’s meaning | A | must be useful for the system, that is, it must serve a purpose in organizing the system’s interaction with the environment. Speaking in terms of evolutionary biology, we may say that snake-detection is what | snake |s were selected for, while ripple-caused | snake |s are fortuitous byproducts (cf. Millikan 1984; Papineau 1987).9

Telexology in a human setting

Natural teleology, in the extended sense of a utility structure, answers to a number of relevant factors, including the specific needs posed by the organism’s body, the complexity and size of the nervous system, the organization of behavior at a social level, and the requirements defined by physical characteristics of the environment. In the case of human beings, we are dealing with cognitive systems that are noted for their enormous complexity and plasticity. They are able to adapt to a wider band of possible conditions, and are better able to adapt to changing conditions, than any other type of system known to us. Moreover, on the theory of neural epistemology outlined in this study, man is engaged in a continual process of literally changing the structure of his brain: new theories mean new ways of responding to reality, involving new neural
structures. In addition, a growing number of external aids enable man to widen his range of cognitive possibilities. Our perceptual system is pros-thetically extended by means of tests, telescopes, and other observation aids, our mnemonic capacities are boosted by libraries and data banks, logical and mathematical skills are enhanced by the use of computers, while the communicative possibilities of human language add a social dimension that affects all of the above aspects of cognition. It is important to realize that all of these more typically human achievements, too, are candidates for teleological analysis. Of each change of neural structure, each new theory, each perceptual aid, the cognitive significance is to be judged in terms of their specific adaptive value. For example, we saw in chapter five that new theories involve a repartitioning of neural vector space. The content of new partitions depends neither on intrinsic features nor on distal causes; rather, it is a question of their functionality in terms of behavioral control, equally engaging both internal and external aspects: how internal structure is used by the organism to selectively respond to new regularities, or to respond to old regularities in new ways.

Social constraints appear to be very much on a par with ecological constraints on content determination. Cognitive performance is calibrated not only by the individual’s bare physical surroundings, but also by various sorts of social structure and dynamics. Habits, practices and institutions, most notably including those of language communities, control the formation of structured responses and behavioral dispositions in individual human beings. Language is obviously of the utmost importance for human knowledge. Indeed, most philosophers of knowledge in the past have seen it fit to base their theories exclusively on the peculiarities of human language, overtly or covertly. I think this is methodologically unwise, however. From the point of view of natural epistemics, language is a behavioral disposition just like others, a set of neural structures specifically calibrated to respond to certain external pressures. Hence, an adequate understanding of language in this purview requires a general theory of how physical and social mechanisms alike may bear on the individual’s performance, the sort of theory offered by teleological accounts.

What is next?
The teleological account of content determination is presently the most promising approach available. Its appeal may be summarized in four points. First, teleology explicitly endorses the externalist framework, which we found to be vastly superior to the internalist alternative. Secondly, the central explanatory
notion of natural function does full justice to the relational nature of cognition, which we found is the key argument for externalism. Thirdly, the teleological solution to the problem of misrepresentation is superior to that of other causal theories. Finally, the notion of a natural teleology is supported by research in a number of other disciplines, including cognitive ethology (Dennett 1987, 237ff), evolutionary biology (Jerison and Jerison 1988; Edelman 1987; Changeux 1985), Gibsonian ‘ecological’ psychology (Gibson 1979; Neisser 1976), and evolutionary epistemology (Campbell 1974). Some of these related theories are admittedly controversial. Yet, the rapprochement noted here is very promising from a naturalistic point of view. There is reason to expect that theories developed in the philosophy of mind, in behavioral sciences, and in natural sciences such as biology, will be able to cooperate to their mutual benefit. The externalist approach outlined here is pre-eminently suited to serve as a general frame of reference for this project, in terms of which the results and conceptual resources of the contributing disciplines may be coordinated.

If teleological theories represent the state of the art in externalism, this does not mean that all problems have been solved. In the next section I consider two residual objections that may be raised against externalist accounts of content, namely, the apparent lack of representational specificity, and the recurring threat of intrinsic content. This last objection will prepare the way for a more ample discussion of transcendental philosophy in the final chapter.

4. Second thoughts about externalism?

What the frog’s eye tells the frog’s brain

The first difficulty with externalist accounts is their apparent lack of representational specificity. The problem is endemic in causal theories. As noted earlier in this chapter, causal theories distinguish between mediated and unmediated chains from object to representation. So far, I simply assumed the identity of the relevant chains to be self-evident. This assumption may be challenged, however. It may be pointed out that causal chains in nature do not come in ready-made chunks. Presumably, they extend indefinitely in time as well as in space. Each cause has prior causes, each effect ulterior effects; moreover, many conditions or parallel chains may bear on the causation of any single effect. How, then, do we decide where the causal chain relevant for the determination of content starts and where it ends? In particular, on what grounds are we justified in taking the distal cause rather than the proximal cause to be the true object of representation? As far as causality is concerned, any link in the causal
chain leading up to a given representation may plausibly be taken to determine content.

The problem envisaged here belongs to a whole family of specificity problems, two other members of which we have already met, namely, the disjunction problem and the \textit{qua}-problem. The \textit{qua}-problem is concerned with \textit{aspectual inspecificity}: how do we decide which aspect of the distal object is being represented by its neural effect? Classical disjunction focuses on \textit{distal inspecificity}: if purported $|A|$s can also be caused by B’s, we may as well say that they mean $|A \ or \ B|$. The present problem, by contrast, is one of \textit{linear inspecificity}: if purported $|A|$s are caused by A’s by virtue of some more distal or more proximal cause P, they may as well be said to mean $|P|$. In what follows, I concentrate in particular on the role of intermediate proximal causes, which strikes me as the most interesting problem here. (For kindred discussions of specificity in causal contexts, see Dretske 1981, ch. 6; Lloyd 1989, chs. 2-3).

The various specificity problems are interrelated in intricate ways. Consider the well-known case of the fly-snapping frog (see fig. 7.4). The frog’s eye-brain-motor structure is such that (given normal neural conditions, which I

![Figure 7.4: What the frog’s eye tells the frog’s brain](image)

Ever since the classic paper on “what the frog’s eye tells the frog’s brain” (Lettvin et al. 1959), frogs and toads have figured prominently in philosophers’ discussions of mental content (typical examples are Dennett 1987; Fodor 1990a and 1990d). The frog in the above figure is tokening a neural representation R, proximally caused by a state of retinal excitation P. Tokens of R cause the frog to snap at the perceived object O, which under optimal conditions is a fly. Intuitively speaking, R’s content is identified as being $|fly|$. Given the fact that also non-fly-caused P tokenings cause the frog to token R, however, causal theories seem to entail that R’s content is better seen as being $|retinal \ excitation \ P|$. In other words, the frog’s eye tells the frog’s brain that it is in a certain state of activation, \textit{not} that it has spotted a fly.
shall assume throughout) certain states of retinal excitation $P$ cause the frog to snap. The required excitation states are typically caused by small, fast-moving objects such as flies, pellets and other ambient black dots. Under environmentally normal conditions the frog’s snapping behavior is conducive to survival: the objects snapped at are usually nutritious, non-toxic (etcetera) flies. Intuitively speaking, we are inclined to identify the content of the frog’s snap-causing representation $R$ as $|fly|$. Yet, the causal evidence is compatible with at least four different ways of assigning content.

2. $R$ means $|fly|$, while its being tokened by pellets is wild. (Similarly for $R$ meaning $|pellet|$.)

3. $R$ signifies the open disjunction $|fly$ or $pellet$ or …$|.$

4. $R$ represents the category $|ambient black dot|$, the aspect of flies and pellets to which $P$ tokens are selectively sensitive.

5. $R$ means $|retinal excitation P|.$

Options (2), (3) and (4) represent externalist strategies for assigning content, while option (5) is typically internalist in the sense discussed in the previous chapter. (3)-(5) may be taken to correspond to three distinct specificity problems. Yet, I think that, in this context at least, they really amount to the same thing. The alleged category of ambient black dots is arguably an open disjunction. It includes objects as dissimilar as flies, elephants, and flagpoles, provided they all project onto the retina under a suitable angle and from a suitable distance. Moreover, the coherence of $|ambient black dot|$s as a category depends entirely on their associated states of retinal excitation. Hence, by definition, $R$ can carry the content $|ambient black dot|$ iff it signals $|excitation P|$. Summarizing these relations, we may say that $|P|$ is the internalist equivalent of $|ambient black dot|$, while the latter is an externalist disjunction in disguise.

Generally speaking, I am inclined to think that these varieties of causal inspecificity always reduce to the same $ur$-problem in this way. I shall not pursue this suggestion here, however. Aspectual and distal inspecificity have been sufficiently discussed above, so I shall confine my attention to linear inspecificity.

Linear inspecificity is apparently resolved by natural teleology. Assuming content to be determined by natural utility, we may argue that the organism’s responding to proximal cause $P$ is useful only by virtue of the fact that $P$ is normally caused by flies. The frog’s $P$-sensitivity was selected by nature for the purpose of fly-detection, not $P$-detection. Fodor (1990b and 1990c) has taken
exception to this solution, however. According to him, the appeal to natural
selection suffers from the same ambiguity as the original causal proposal. In
the following passage, for example, taken from a discussion of the disjunction
problem, Fodor argues that it is equally useful for frogs to respond to flies
than for them to respond to ambient black dots.

“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 mechanisms
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. “Erst kommt das Fressen, dann kommt die
Moral”. Darwin cares how many flies you eat, but not what description you eat
them under. (...) I think that the right answer is that appeals to mechanism
of selection won’t decide between reliably equivalent content ascriptions; i.e.,
they won’t decide between any pair of equivalent content ascriptions
where the equivalence is counterfactual supporting” (Fodor 1990b, 72-73,
italics in original; German corrected, JS).

The same line of reasoning applies to retinal states P, which are by definition
caused by ambient black dots. |ambient black dot| and |retinal state P| are
reliably equivalent; the mathematics of survival come out the same on either
description.10

Fodor’s objection calls for a number of comments. In the first place, there is
a well-known problem with counterfactual supporting coextensive predicates.
The classic example, due to Nelson Goodman (1953, 59ff), involves the predi-
cate ‘grue’, which applies to all things examined before some future time t just
in case they are green, but to all other things just in case they are blue. Under
present conditions, the contents |green| and |grue| are reliably equivalent: it
is impossible to decide whether it is the object’s greenness or grueness that
causes me to token R. As Kim Sterelny (1990, 126-127) has noted, part of
Fodor’s problem is based on this ‘new riddle of induction’. In the case of
gruness and greenness, we are intuitively certain that greenness is doing the
relevant causal work. Similarly, if |dot|s, |fly|s, and |P|s are under present
conditions reliably equivalent, we may still be confident that it is flies that are
doing the relevant work: they are what nature has selected the frog’s dot-P-R
mechanism for. The intuitions involved are admittedly difficult to capture. Yet,
it is important to realize that Goodman’s riddle is not peculiar to the present
account of mental content: it is a standard problem for any theory of knowledge.

One way to try to explicate the relevant intuitions in the example is by invoking Fodor’s notion of asymmetric dependencies. As transpires from the above quotation, Fodor seems to agree that the dot-to-R connection would not have been selected by evolution if not for the fly-to-R connection holding: dots would not have caused snapplings if they were not normally flies. The reverse does not hold, however: evolution would favor frogs snapping at flies even if flies were not dots. In short, frogs snapping at dots is asymmetrically dependent on their snapping at flies. Similarly for the P-to-R and fly-to-P connections: speaking in terms of natural selection, P’s causing snaps depends on flies causing P’s, but not the other way around. Evolution might favor frogs tokening P’s in response to flies, even if dots did not cause snaps, as, for example, in environments where many dots are non-edible objects that merely look like flies. In short, dots causing snaps is asymmetrically dependent on flies being dots: if dots are snapped at they must normally be flies, while if dots are not snapped at they may (but need not) be flies.

These considerations sufficiently demonstrate that, pace Fodor, natural selection is not at all indifferent with regard to contextually equivalent descriptions. Heeding the relevant counterfactuals, it is perfectly clear what a given neural structure is selected for. Hence, if content is determined by functionality, as the teleological theory has it, then R must specifically mean |fly| (cf. Millikan 1991; Dennett 1987, 314ff).

There is a residual worry, however, which indeed seems to be the basic motive behind the black dot argument. Fodor is worried that functionality is in the eye of the beholder. He thinks this is a problem because content should be in the eye of the frog, so to speak. We may be able to see through the teleological structure of the frog’s brain, but the frog does not see things this way. As far as he is concerned, ambient black dots are all he has to go on, and they are just fine; as Fodor puts it, they are “simply super” (op. cit., 75). The implicit idea is that the frog’s brain relies entirely on what the frog’s eye tells it. It cannot see the world beyond the retina: if the eye says |P|, then the brain knows there are ambient black dots to be snapped at, and that is the end of the story. This line of reasoning is familiar enough from the previous chapter. Like so many other arguments for internalism, it rests on the mistaken assumption that cognitive science should be conducted from the point of view of the Cartesian frog, the ‘internal subject’ that is locked outside the world. I refer back to the objections raised against this view in the previous chapter: cogni-
tion is definitely *not* a matter of gazing at the back of one’s retina. Boldly put, it is being-in-the-world.

**Remotely controlled brains?**

The above reflections bring me to a second difficulty with externalism, the recurring threat of intrinsic content. On the face of it, external theories seem to be far removed from the concerns of Cartesianism. Privileged, internally defined mental contents are explicitly rejected; organisms are treated as *being* subjects rather than as *containing* them; a strongly biological bias precludes all metaphysical dualism. Still, we now face a danger that is the exact opposite of internalism but arguably no less Cartesian in nature, namely, that organisms are reduced to hollow automata, remotely controlled by the surrounding world. This is of course Descartes’ well-known view of the brutes and of the human body, mechanical devices which as parts of the *res extensa* are ruled by mechanical laws. When the *res cogitans* is removed, what is left of man is a physical system on a par with clocks, computers, and door bells. Now, if we continue to look upon the brain as the seat of control where behavior is instigated, we find that the brain itself is remotely controlled by the environment. We have become the puppets of the world: if our tokening R means \( \text{apple juice} \), this is because the world’s apple juice made it do so.

The remote control caricature of mental content dramatizes the fact that, according to externalism, individual subjects are not the *source* of their so-called intentionality. I think this rejection of internalist philosophy is perfectly correct. However, by magnifying the role of the environment, externalist theories are prone to reintroduce intrinsic content. Meanings may not be inside the head, but apparently they are still ‘given’ in the sense of being fixed by inalienable links between symbol and content. If internal psychology does not define these links, then biology will. Hence, intrinsic features of the world take the place of features intrinsic to the individual; one way or the other, content will be intrinsic.

Paradoxically enough, then, the ‘remote control’ variety of externalism remains firmly wed to the Myth of the Given, even if there is no longer an internal subject to read off the ‘givens’. Representations are still credited with having a proper epistemic identity, which is dictated by nature itself. By the same token, externalism may be charged with committing the naturalistic fallacy. Surely, the justification of our tokening \( \text{A} \) cannot be that it was caused by the world’s intrinsic A-ness. Any attempt to understand knowledge along these lines strikes us as simply spurious.
The problem noted here shows that it is very difficult to strike a clear balance between the two poles of cognition, objective and subjective. In the concluding chapter I present this problem from a broader historical and systematic perspective. As we shall see, there is a strong temptation in transcendental philosophy and psychology to analyze knowledge, as a relation between subject and object, in terms of either subject or object, but not both at the same time. I think this temptation can and should be withstood, however, and that the teleological approach holds the best chances for doing so. As I have tried to explain in this chapter, teleological theories call for a drastic reorientation with regard to the nature of content. They offer a radical alternative to all traditional accounts that ground content in intrinsic features of either subject-as-such or object-as-such. The strength of teleology is the fact that it fully respects the relational nature of cognition (see, for example, my discussion of epistemic privilege and functional misrepresentation). Natural function and utility, the basic teleological constraints on content determination, are eminently suited for this purpose, focusing on the interaction between organisms and their environment rather than on neural make-up or ecological lay-out alone. Succinctly put, flies do not cause \textit{fly}s because they are intrinsically flies, but because they are useful-for-frogs (which is certainly not an intrinsic feature of either flies or \textit{fly}s, let alone of frogs). Unlike other species of externalism, then, the teleological account outlined here may successfully withstand the remote control imputation, and avoid the extravaganza of intrinsic content.
Chapter eight
Real knowledge in perspective

In this final chapter I want to draw together the main strands of argument presented in this study, and put them in a broader historical perspective. The ongoing debate on mental content repeats issues that have been discussed in philosophy for over two thousand years. I think we will be better able to appreciate some of the points made in this study against the backdrop of their historical development.

Section 1 is a short history of knowledge from Aristotle to Kant. Internalism and externalism are identified as the two main determining vectors in the force field of traditional philosophy. Stripped to their conceptual bone, they lead us to expect that no solution composed of internalism and externalism will ever be able to deal with knowledge in a satisfactory way. A radical alternative is needed.

I turn to such an alternative in section 2. Taking my lead from a generalized notion of a transcendental deduction, I draw the outlines of a new framework for cognitive research, called ‘relationism’. Unlike internalism and externalism, this new framework can meet the requirements of the dual aspect theory of knowledge introduced in chapter three: that a feasible account of knowledge must respect both its epistemic and its natural aspect. One promising way to work relationism into an empirical theory of cognition is natural teleology, as described in the previous chapter. I conclude this section with some reflections on the connection between relationism and Putnam’s ‘internal realism’.

Section 3 is an examination of how the relationist alternative bears on the theory of content. Taking as my example the controversy on Gibsonian psychology, I argue once more that we are neither remotely controlled by the environment, nor endowed with intrinsic contents. Then it is time to sum up and look ahead: in the final section I canvass the prospects of cognitive science from a relationist point of view.
1. A short history of knowledge

The issue of internalism and externalism in cognitive science reflects a long-standing controversy in the history of philosophy. From Antiquity up to the present day, philosophers have asked themselves what determines the content of our ideas. Historically speaking, there has been a shift from a predominantly externalist outlook in Antiquity and the Middle Ages, exemplified by Aristotelian-Scholastic psychology, to the internalist philosophies of Descartes and Kant. This development coincides with a shift from metaphysics to epistemology as the primary discipline or ‘first philosophy’. It is instructive to compare the modern positions on content determination with their classical counterparts: the historical perspective facilitates our understanding of the conceptual backbone of the positions involved. In this section I discuss three significant episodes in the history of knowledge: Aristotelian psychology, the rise of Cartesian dualism in the seventeenth-century, and Kant’s project of pure reason. My reconstruction of these views will be fairly schematic. Bear in mind that my ultimate aim is conceptual rather than historical. (For more detailed studies, see, for example, Meijering 1989a; Spruit 1994.)

Aristotelian-Scholastic psychology and semantics

Aristotle’s theory of human psychology was essentially a chapter in natural philosophy. Especially as it was received in Scholastic thought, his metaphysics of nature pivoted on the doctrine of hylomorphism, the theory that all (created) being consists of two aspectual components, matter (potentiality) and form (actuality). Form is what actualizes the potentiality of matter, matter is what individualizes universal form. Simply put, trees are wood with the form of life, stools are wood with the stool-form; wood is earth with the wood-form, stone is earth with the stone-form. The form of an object contains its intelligible properties, including its essence. In the case of human beings, for example, the soul (ψυχή, anima) acts as the form of the body. It actualizes the slumbering potentialities of the material substrate, and it contains the essence of man, which is that of a rational living being (animal rationale). Other living things are less perfect than man: their souls lack the specifically human addition to the form of life that is reason (see figure 8.1).

Also specific psychological functions, such as perception and cognition, are explained in terms of hylomorphism. For example, a sleeping man is to a waking man as matter is to form: when awake, we actualize the potentiali-
ties inherent in our soul. The nature of the human soul is such that it is able to receive the forms of external objects. Perceptual knowledge consists in the absorption of the form of the objects perceived: when the object impinges on a sensory organ via an appropriate medium, the subject is literally in-form-ed by it, becoming qualitatively identical to the object.

“In general, with regard to all sense-perception, we must take it that the sense is that which can receive perceptible forms without their matter, as wax receives the imprint of the ring without the iron or gold” (Aristotle, *De anima*, 424a17, transl. Hamlyn).

In the case of vision, for example, color and shape are transmitted from the object to the eye, which literally takes on the form of the object and becomes qualitatively identical to it. The chain of forms may then extend further inside the subject: the form received by the eye is passed on to the common

Figure 8.1: The structure of the soul in Aristotelian-Scholastic psychology
In Aristotelian psychology, the structure of the soul (ψυχή, *anima*) is accounted for in terms of hylomorphism. Generally speaking, the soul is defined as the form of the bodily substrate. Actualizing powers inherent in the organic body, the soul manifests itself in the form of specific functions such as nutrition, locomotion, and intellect. (In this respect, Aristotelian psychology is a precursor of functionalism in the philosophy of mind; see, for example, Putnam 1975a, 291ff; Cohen 1992; this interpretation is disputed, however, by Heinaman 1990.) The functions of the soul may be ordered hierarchically in order of perfection. Thus, the vegetative soul is responsible for nutrition, growth and generation. The sensitive soul, which is to the former as form is to matter, is responsible for perception, appetite and locomotion. The rational soul, which is specifically human, is a perfection of the former, and is responsible for intellect and volition.
Real knowledge in perspective 241

sense (*sensus communis*), where the information from different senses is fused to form an overall picture of the object’s intelligible aspects. Ultimately, from this super-percept the essential properties are extracted by the intellect, resulting in intellectual knowledge of the object’s essence.

To illustrate the point of this theory, consider its bearing on classical and medieval semantics. Figure 8.2 pictures the theory of meaning for natural language as proposed by Aristotle in *De interpretatione*, and as received in medieval philosophy through Boethius’ commentary on this work.² In classical semantics, written expressions are meaningful because they relate to spoken words, and spoken words are meaningful because they express thoughts. These relations rely on man-made conventions: different languages may use different signs, both written and spoken, to express the same thought. The meaning of our thoughts, by contrast, is not determined conventionally but *naturally*. Thought is meaningful by virtue of its relation to reality. Our ideas mean definite objects and properties in the world because

![Diagram](image)

**Figure 8.2: Classical semantics**

Semantics of natural language and of mind according to Aristotle (*De interpretatione*) and Boethius (*In de interpretatione*). Written expressions (*litterae*) primarily signify spoken sounds (*voices*), which in turn signify mental contents (*intellectus*), or affections of the soul (*passiones animae*). Mental contents are natural signs of objects (*res*), by virtue of the similarity (*similitudo*) between content and object. This is a natural relation (*natura*), as opposed to the conventional relations (*positione*) between writing, sound, and thought.
they are resemblances (*similitudines*) of these objects and properties. Metaphysically speaking, as we have just seen, concepts and other affections of the soul literally have the same *form* as the objects they represent.

In terms of the previous two chapters, the Aristotelian-Scholastic theory of cognition is a form of externalism *par excellence*. Mental content is determined purely by factors external to the individual subject. I draw attention to two premises on which this position is based. The first is that of metaphysical realism: the world is endowed with an intelligible structure of its own. In Putnam’s phrase, this position rests on the assumption of a “ready-made world” (Putnam 1981; 1983, 205ff). Secondly, the subject is claimed to be able to receive all forms, and to simply grasp them without substantial alteration or distortion. The chain of forms from object to intellect is cognitively homogeneous: apart from abstraction, no real internal processing is required.

Notice, incidentally, that the metaphysics underlying this position was by no means a gratuitous assumption. For many centuries, the hylomorphic scheme served as the basis of a highly successful, comprehensive worldview, which offered coherent explanations for a wide range of phenomena.

*Cartesian metaphysics and its skeptical implications*

A radically new theory of perception and cognition was required by the metaphysical change from hylomorphism to corpuscularism in the sixteenth and seventeenth century. Aristotle’s ontology of forms was replaced by a vortex of minuscule particles acting in strictly mechanical fashion. Psychologically speaking, it is no longer possible to think of perception as the direct reception of forms by the soul. Billions of formless physical stimuli now bombard the body. It is not at all clear how, under these circumstances, knowledge is possible at all (Meijering 1989a, chs. 1 and 6).

The skeptical implications of corpuscularism fostered the epistemological concern that is typical of modern philosophy. Implicit in this reorientation is a development towards internalism, as is particularly clear in the work of Descartes. I mention three aspects of this development that are most pertinent in this context—one epistemological, one metaphysical, and one psychological.

In the first place, corpuscularism opened up an epistemological gap between the ‘manifest image’ and the ‘scientific image’ of reality (cf. Sellars 1963, ch. 1). Apparently, the world is not what it seems to be. Hence, the reliability of sensory evidence needs to be justified (as must, paradoxically,
the new science to be founded on it). Descartes is famous for developing this insight into his project of ‘methodical doubt’. In a well-known series of arguments presented in the *Meditations*, he considered the possibility that our knowledge of the external world is radically illusory. Yet, even on the extreme condition that the external world does not exist, our ideas about objects would retain their meaning, and would be given to the mind with indubitable force. In the new Cartesian framework, givenness-to-the-mind becomes the hallmark of mental content, a doctrine most poignantly expressed in the theory of innate ideas. As we have seen in earlier chapters, this aspect of Descartes’ philosophy survives in the widespread presumption that content must be determined intrinsically. Notice, incidentally, that Descartes’ claim was strictly speaking unthinkable in Aristotelian psychology: without objects to in-form us, no meaningful contents can be formed.

Secondly, Descartes’ “invention of the mind” (Rorty 1979) opened up a metaphysical divide between subject and world. In Cartesian philosophy, the *res cogitans* that is the thinking subject was posited as a separate substance next to the *res extensa* that is the physical universe. An obvious problem for this position is how body and mind are able to interact: if they have nothing in common, being separate substances, how can the mind cause changes in the body, or vice versa? Descartes’ solution was to locate a point of contact between body and mind in the pineal gland (thereby violating his own metaphysical principles). Two remarkable consequences of Descartes’ dualism should be mentioned here. On the one hand, by assigning matter and mind to separate realms of reality, Descartes effectively removed the subject from the world. One consequence of this move, which survives in modern cognitive science, is a tendency to deny the physical reality of knowledge. On the other hand, by locating a point of contact inside the brain, Descartes effectively imprisoned the subject inside the body, condemning him to read only the mechanical symbols written on the pineal gland, and to interact with the world only indirectly. This claim has been identified as the Myth of the Cartesian Theatre (Dennett 1992; Ryle 1949): the idea that there is a place inside the brain where all sensory information comes together for inspection by the mind, and where all bodily action is instigated.

Finally, a tremendous psychological problem was posed by the fact that the nature of the physical stimulus, according to the new theory, is totally different from that of the mental response. Instead of qualitative copies of distal objects, a puzzling variety of mechanical stimuli is presented to the mind. How does the mind know how to translate these signals into the ap-
propriate ideas? Aristotelian signals wore their ideas on their sleeves, but this option is no longer available. Seventeenth-century philosophers generally believed that the stimulus still contains sufficient information about its distal cause, but that it needs to be decoded first. A theory of internal processing should account for the mind’s ability to reconstruct the nature of the distal cause from their effects on the body. Descartes, for example, suggested that ideas are ‘ordained by nature’ to follow upon certain movements in the pineal gland. When the pineal gland moves in an appropriate manner (signaling the presence of a cow, for example), the mind is magically ‘ordained’ to token the concept \textit{cow}. Similar solutions can be found in other rationalists such as Malebranche, Spinoza and Leibniz. What binds these positions together is their allegiance to a quasi-magical correspondence (a \textit{Präformationssystem}, as Kant called it; 1781, B 167) between ideas and the world.

It is instructive to compare Descartes’ solution to that of Locke. Both Descartes and Locke made a sharp distinction between primary and secondary qualities of objects. Unlike Aristotelian forms, secondary qualities do not inhere in external objects as such. Rather, they are effects caused in the perceiver by the object. Examples of secondary qualities include smell, sound, coldness and warmth, the perception of which is typically dependent on the nature of the subject. Primary qualities, by contrast, are assumed to be truly independent of the subject, and to inhere in the object as such. They include geometrical and arithmetical properties: an object’s being \textit{triangular}, or its being numerically \textit{one}, for example, are properties that do not depend on the nature of the perceiver. Rather, they reflect the essential ontology of the physical world as it is in itself. How these primary qualities are perceived is explained by Locke in a famous passage of the \textit{Essay concerning human understanding}:

“For the Objects of our Senses do, many of them, obtrude their particular Ideas upon our minds whether we will or not (…). These simple Ideas, when offered to the mind, the Understanding can no more refuse to have, nor alter them when they are imprinted, nor blot them out and make new ones itself, than a mirror can refuse, alter or obliterate the Images or Ideas which the Objects set before it do therein produce. As the Bodies that surround us do diversely affect our Organs, the mind is forced to receive the Impressions; and cannot avoid the Perception of those Ideas that are annexed to them” (Locke 1670, II, 1, § 25).
The ideas corresponding to primary qualities are simply ‘obtruded’ to the mind ‘whether we will or not’. From these ‘simple ideas’, the mind then reconstructs the world as we know it. Locke’s project for philosophy was to develop a sort of mechanics of the mind, “eine gewisse Physiologie des menschlichen Geistes”, as it was called by Kant (1781, A x). This mental companion to Newton’s mechanics of matter was meant to account for the construction of concepts and judgements from the original sense impressions.\(^3\)

Locke’s outlook may in many ways have been different from Descartes’, yet his ‘solution’ to the psychological problem is remarkably similar (cf. Adams 1975). Both Descartes and Locke believed that the world has a subject-independent, ontological structure, determined by mathematical properties (\textit{extensio}). This structure is communicated to the mind, but the mechanism responsible for the transition remains obscure: the right ideas simply pop up at the right moment.

\textit{Kant’s radical project of transcendental philosophy}

If Descartes gave modern philosophy its characteristically internalist turn, it was Kant who took internalism to its logical conclusion. The Cartesian framework still incorporated important elements of Aristotelian philosophy. Thus, it retained the idea that the external world has a metaphysical structure of its own, which can be specified mind-independently. Moreover, knowledge remained in the final analysis a formal correspondence between mental content and world. Kant’s contribution was to remove these residual elements of externalism. Several thinkers prepared the way for Kant. I mention two of them here, Berkeley and Hume.

Working within the internalist framework set out by Descartes, Berkeley showed that primary qualities in a non-Aristotelian setting are simply impossible (cf. Putnam 1981, 58ff). \textit{All} qualities of objects are subject-dependent; in other words, all qualities are \textit{secondary}. Consider first our perception of geometrical properties. So-called rectangular objects, such as books, are rarely perceived as pure rectangles; under most conditions of perception, they appear to us in a variety of rhomboid shapes. The decision to \textit{think} of books as being rectangular is a construction on the part of the subject, involving an elaborate scheme of geometrical transformations. Much the same argument holds for arithmetical properties, which also depend on conditions of perception. Whether an object, or a collection of objects, is perceived as \textit{one} or as \textit{many}, depends on the distance at which it is perceived: what appears as numerically \textit{one} from a distance, may appear as \textit{many} at close quarters (for
further discussion, see Schwartz 1994).

An objection that naturally suggests itself here is that some conditions of perception are more appropriate than others. Surely, it may be argued, the true number of objects can be verified even by a blind man: let him approach to a suitable distance and simply count the objects with his hands. This solution will not do, however. Berkeley’s argument holds for other sensory faculties as well as for vision. With regard to touch, for example, what appears to be numerically one when touched by adult human beings, may appear as many when touched by smaller hands. What would a creature the size of an amoeba make of the apparent unity of my armchair? Similarly, if Mount Blanc appears to me a solid unity, which I can climb and touch and fly around in an airplane, what sense does this make to a creature the size of a galaxy, with suns for fingers and stardust for feet? (Notice, incidentally, that this line of reasoning points toward a teleological approach to perceptual conditions, as discussed in the previous chapter.)

In an even more skeptical vein, Hume argued that what is given to the mind is only a flow of disparate impressions; neither substantial unity nor causal necessity can be found therein. Therefore, if we think of an apple as a body having certain properties, or if we think of a causal nexus between two events, we go beyond the evidence that is given to the mind. We never perceive any structure in the world; we simply bundle impressions in certain ways. Hence, our idea of the world as a structured ensemble of interacting bodies is largely a construction on the part of the subject; of the real structure underlying our impressions we are necessarily ignorant.

Kant’s Critique of pure reason was the endpoint of this development. While Aristotle’s object-oriented approach to knowledge had partially survived the turn to Cartesian internalism, Kant suggested a more radically subject-oriented approach, a reorientation which he famously compared to the Copernican revolution in astronomy.

“Hitherto it has been assumed that all our knowledge must conform to objects. But all attempts to extend our knowledge of objects by establishing something in regard to them a priori, by means of concepts, have, on this assumption, ended in failure. We must therefore make trial whether we may not have more success in the tasks of metaphysics, if we suppose that objects must conform to our knowledge. (…) We should then be proceeding precisely on the lines of Copernicus’ primary hypothesis. Failing of satisfactory progress in explaining the movements of the heavenly
bodies on the supposition that they all revolved round the spectator, he tried whether he might not have better success if he made the spectator to revolve and the stars to remain at rest” (Kant 1781, B xvi, transl. Kemp Smith).

According to Kant, our understanding of knowledge should not be based on the constitution of the object, but on that of the subject. It is the subject’s internal cognitive make-up that determines the nature, scope, and validity of knowledge. Henceforth, epistemology definitely takes priority over metaphysics as first philosophy.

Taking his lead from Hume’s theory of knowledge, Kant argued that whatever order we perceive in the world is really a product of our cognitive apparatus. As Hume had shown, neither necessity nor universality enter the mind from without. What is given from without is a mere multitude of disparate stimuli. Science, however, deals with necessary and universal judg-

Figure 8.3: Kant’s transcendental analysis of knowledge
In the faculty of sensibility, the pure manifold of impressions is ordered in space and in time, producing an intuitive representation of the object. Intuitions may be given conceptual content by the understanding, which consists of a matrix of twelve basic categories. Notice that the structure of knowledge, according to Kant, rests on a two-layered distinction between form and matter: intuitions consist of matter (the raw material of the manifold) and form (space and time), while concepts are to intuitions as form is to matter.
ments. Hence, if science is to be possible at all, the structural aspects of knowledge must be imposed by the mind. This is Kant’s famous transcendental argument: the possibility of scientific knowledge requires that the subject be structured in definite ways. By determining the perceptual and conceptual schemes implied by our knowledge of the world, including in particular Euclidean geometry and Newtonian physics, Kant deduced the nature of the internal structure responsible for this knowledge.

Figure 8.3 shows part of the relevant structure. When the faculty of sensibility is affected by an object, the first order that is imposed on the incoming signals is that of space and time. Space and time are not properties of the world as it is in itself, but formal aspects of human perception, or ‘forms of intuition’. When the raw material is thus ordered, the faculty of sensibility has produced an intuitive (non-conceptual) representation of the object (Anschauung). Next, conceptual content is imposed on these intuitions by the faculty of understanding, which consists of a matrix of twelve basic categories. Examples of these ‘pure concepts’ are |unity|, |existence|, |substance| and |causality|. None of the categories represents properties of the world as it is in itself; rather, the categories are just various modes of processing information. Which concepts to apply to which set of intuitions is determined by a set of rules or schemata, which link the structure of sensibility to that of the understanding. The schema for substance, for example, is “permanence of the real in time”; hence, the concept |substance| is applied to that in the intuition which stays the same when other things change. Similarly, the concept |cause| is applied to “the real upon which, whenever posited, something else always follows” (op. cit., A 137ff, B 176ff).

The intrinsic structure of the subject determines the phenomenal world, that is, the world as it appears to us. The noumenal world, by contrast, the world as it is in itself, is beyond all perception and conceptualization. At this point Kant’s philosophy proves particularly uncompromising. We would like to be able to explain how the world as it is in itself constrains our empirical knowledge, but on Kant’s account this is impossible by definition. Because all order is imposed by the subject, the object as such is utterly ineffable. The Ding an sich is strictly a noumenon—the abstract idea of an object of knowledge that remains when all our knowledge of it is bracketed. We cannot say that it is a substance, that it is one or that it is many, nor even that it exists, or that it is a thing. For all these statements refer to specific aspects of the order imposed by the subject itself.

Much the same goes for the ‘stimuli’ (Affizierungen) received by the sub-
ject. In earlier internalist philosophies, stimuli were assumed to carry enough information about their distal source to determine how they should be processed. There remained a formal correspondence between object, stimulus and concept. On Kant’s account, however, the stimulus is radically ineffable. Not only is it impossible to explain its action on the mind as a causal process (because causality is a category imposed by the mind), but it is also impossible to say that it carries information, or that it has a cognitive content of its own. The stimulus as it affects the mind is, in Kant’s words, a ‘pure manifold’ (reine Mannigfaltigkeit): it is not a unity nor a plurality (which are concepts of the understanding), it is ordered neither in space nor in time (which are forms of the intuition), it is simply a ‘form-less’ X beyond all cognitive content. Whatever mental representations are eventually occasioned by the stimulus, their content derives entirely from the subject’s intrinsic structure, not from some external source.

From ontological hylomorphism to epistemological hylomorphism

A convenient way to summarize the development sketched in this section is to concentrate on the role of hylomorphism. Interestingly, the matter/form distinction of Aristotelian philosophy is redeployed in post-Cartesian philosophy, where it is transposed from the object to the subject (see figure 8.4).

Figure 8.4: Metaphysical and epistemological hylomorphism

In modern post-Cartesian philosophy, the Aristotelian metaphysics of matter and form (left) is transposed to the subject-part of cognition, where we may speak of an ‘epistemological hylomorphism’ (right). In contemporary philosophy of science, the dualism of form and matter survives as the dualism of scheme and content: conceptual schemes are assumed to impose structure on the raw material delivered by observation. For some critical studies of this idea, see Quine 1953; Davidson 1974; Rorty 1972 and 1979; Rescher 1980; Mary McGinn 1982. In an intelligent discussion of the ‘linguistic consensus’ in modern philosophy, Pols (1992, 67ff) speaks of two ‘dogmas’: that of ‘formative rationality’ or ‘linguistic enclosure’ (form), and that of the ineffable empirical stimulus (content).
The new hylomorphism is strictly epistemological, as opposed to the metaphysical hylomorphism of Aristotelians. This is particularly clear in the work of Kant. If we compare his position to that of Aristotle, several differences spring to mind. First, matter and form are aspects of the subject, not of the object. On Kant’s account, ‘form’ is the cognitive structure of the subject, while ‘matter’ is the pure manifold of affections supplied by the object. Moreover, knowledge is analyzed not as a transfer of forms, but as a transfer of matter. Finally, it is the object rather than the subject that is ‘transcendently indeterminate’, that is, able to receive all forms.

I think of the two positions schematically reconstructed here as two abstract vectors in the force field of philosophy. (Recall my remarks on categorical analysis in chapter two.) Whether or not any one thinker actually endorsed either of these ‘pure’ positions is less important than the fact that they represent conceptual forces pulling in different directions: internalism and externalism. In point of fact, many philosophers took mixed positions, represented by composite vectors. As we have seen, this is certainly true of the positions of Descartes and Locke.⁴

In this study I have pressed the need for an approach to knowledge that respects both the epistemic and the natural aspect of cognitive states. The history of knowledge as reconstructed here strongly suggests that this requirement will never be met as long as we continue to frame our theories in terms of pure internalism or pure externalism. The two vectors simply pull in different directions: internalism pulls toward an approach that is epistemic but not naturalistic, while externalism pulls toward an approach that is naturalistic but not epistemic. Aristotle, for example, argued that perception and cognition are natural processes, but was unable to account for the specifically cognitive aspects of these processes. Kant, on the other hand, argued that cognition must be studied from the point of view of the ‘transcendental’ subject, making knowledge epistemic but not real.

If my analysis is correct, no solution can be found in the plane of internalism and externalism itself. No position composed of internalism and externalism will be able to meet the necessary requirements. Breaking the deadlock calls for something more drastic, something beyond internalism and externalism. In the next section I turn to such a radical alternative. Speaking in terms of the vector image, I want to consider adding a new dimension to the problem field, orthogonal to the plane of internalism and externalism.
2. Transcendental philosophy reconsidered

In a seminal paper on artificial intelligence, Dennett describes the aims of traditional philosophy as being continuous with those of AI and cognitive psychology (Dennett 1978b).

“Faced with the practical impossibility of answering the empirical questions of psychology by brute inspection (how in fact does the nervous system accomplish X or Y or Z?), psychologists ask themselves an easier preliminary question: How could any system (with features A, B, C, ...) possibly accomplish X? This sort of question is easier because it is ‘less empirical’; it is an engineering question, a quest for a solution (any solution) rather than a discovery” (Dennett 1978b, 110-111).

The ‘engineering question’ raised here is in fact a species of the transcendental question asked by Kant, ‘How is knowledge possible?’ The only difference between Kant and psychologists lies in the choice of empirical constraints on the blanks in the question schema.

“The more empirical constraints one puts on the description of the requisite behavior, the greater the claim to ‘psychological reality’ one’s answer must take. For instance, one can ask how any neuronal network with such-and-such physical features could possibly accomplish human color discriminations (…). Or, one can ask, with Kant, how anything at all could possibly experience or know anything at all. Pure epistemology thus viewed, for instance, is simply the limiting case of the psychologist’s quest” (op. cit., 111).

I think that Dennett’s suggestion here is essentially correct. As we have seen in the previous section, basically the same questions as raised by modern psychology have been asked by philosophers throughout the ages. What is more, basically the same types of answer were given then that are now mooted in the theory of content, namely, internalism and externalism. It is significant that Dennett, in the paper quoted here, considers only the internalist side of cognitive research: the constraints that are added to the transcendental question schema are typical examples of internalist features, such as neural make-up, or narrowly construed behavioral capacities like color discrimination. I am not saying that internalist constraints are not important,
or that this type of research has no value at all. Yet, the widespread presumption in favor of internalism should not blind us to the fact that externalism is an equally valid approach to answering the same question: how is knowledge possible? Or, in more distinctly externalist terms, how must the world be structured for us to be able to know it?

A generalized notion of transcendental deduction

Like Dennett in the passage above, and like many other philosophers of cognitive science, Kant considered the problem of knowledge from a characteristically internalist perspective. When he devised his idea of a transcendental deduction, he was thinking of establishing the internal structure of the cognitive subject. I want to generalize on this idea, and to consider the question of transcendental conditions from a more abstract point of view.

The aim of a transcendental deduction is to establish the necessary and sufficient conditions for something to be possible. If the object of inquiry is knowledge, a number of options is available. In the Western tradition, ‘knowledge’ is generally conceived as a relation between subject and object. In and by itself, this concept is perfectly neutral with respect to the various metaphysical and epistemological accounts that have been given of it. Applying the idea of a transcendental deduction from this abstract angle, the two traditional approaches to cognition can be identified as two different ways of locating transcendental conditions, namely, as lying either at the subject side of cognition, or at the object side. The definitions of internalism and externalism given in chapter six may now be restated as follows:

1. **Internalism**: the conditions of knowledge are located in the individual subject. Knowledge is possible because the subject has a capacity for knowledge. A corollary of this view is that the object is transcendentally indeterminate.

2. **Externalism**: the conditions of knowledge are located in the external world. Knowledge is possible because the world is intelligible. A corollary of this view is that the subject is transcendentally indeterminate.

Notice that these definitions are more radical than their counterparts in chapter six. Especially the definition of externalism, which in chapter six implicitly allowed for partial reference to internal factors, has been sharpened. Definitions (1) and (2) may be seen as the ‘pure vectors’ of which the positions described in chapters six and seven are composed.
The claims made in (1) and (2) are mutually inconsistent. Moreover, they are both stated in fully circular terms. This circularity is not something to be worried about, however. It is inherent to taking a transcendental perspective on knowledge. As Dennett put it in the above quotation, transcendental theories are solutions, not discoveries: their purpose is conceptual, not empirical. They are not put forward as an explanation of knowledge (which would make them viciously circular), but rather as a framework within which to conduct the empirical study of knowledge. Obviously, the frameworks defined by (1) and (2) are very different from one another: as we have seen in chapters six and seven, they set out two very different lines of empirical research.

Stripped to their bare conceptual bone, the traditional frameworks for studying knowledge reveal themselves as fundamentally implausible. They miss the mark by tending to absolutize the contribution of either the subject-as-such or the object-as-such. Considering the fact that knowledge is first of all a relation between subject and object, it is clear that no such approach will ever be completely successful.

At this point, however, an easily overlooked alternative suggests itself, which I shall call ‘relationism’:

3. Relationism: the conditions of knowledge can be found neither in the object-as-such nor in the subject-as-such. Rather, the conditions of knowledge reside essentially in the relation between subject and object. Knowledge is possible because the cognitive structure of the subject fits the intelligible structure of reality. Conversely, the structure of the object fits that of the subject. On this view, neither subject nor object are transcendentally indeterminate.

As a transcendental claim about the nature of knowledge, relationism is again stated in fully circular terms. Still, it describes a conceptual framework for cognitive science that is radically different from internalism and externalism. It invites us to search not for the intrinsic form of the object-as-such, nor for the intrinsic processing structure of the subject-as-such. It reminds us to keep an open mind for constraints on both sides of the knowledge relation, and to understand these in terms of their mutual relation.

One particularly promising way to work this idea into a full-fledged theory is that of natural teleology. As explained in chapter seven, the theory of natural teleology thinks of knowledge in terms of a utility structure, rather than in terms of neural make-up or ecological layout alone. Focusing on the
interaction between the organism and its environment, it understands the organism’s internal representations as being useful for this interaction. The natural utility structure comprises a number of relevant factors, including the specific needs posed by the organism’s body, the complexity and size of the nervous system, the organization of behavior at a social level, and the requirements defined by physical characteristics of the environment.

Relationism is located in a plane orthogonal to that of internalism and externalism. Its solution to the problem of knowledge is not a mixture of internalism and externalism, although it combines insights of both. To illustrate this point, consider the difference between relationism and the hybrid form of internalism endorsed by Descartes. While Descartes officially rejected Aristotelian hylomorphism, he retained the idea of a formal correspondence between mind and world. At first sight, this ‘fit’ between cognitive structure and real structure may seem a truly relational solution. Unlike the relationism proposed here, however, Descartes’ dualism has in principle no room for explaining the interaction between mind and world. The auxiliary hypothesis of the pineal gland is inconsistent with his general metaphysics, and the formal correspondence remains virtually miraculous. The deeper reason for this incongruity is the fact that Descartes simply grafted his new internalism onto the externalist heritage. Relationism, by contrast, takes a different stand, allowing us to see the object and the subject as inhabitants of the same natural world. The subject is no longer locked up inside the body and outside the world; rather, it is a body, and it is in the world. By this token, the interaction between subject and object is readily explained as a natural process without any essential discontinuity; there is no need to close a ‘gap’ between ordo essendi and ordo cognoscendi. At the same time, however, relationism fully respects the epistemic aspect of knowledge, namely, by identifying cognitive states (propositional attitudes) as relational states of individual subjects.

Relationism, anyone?

Over the past ten or fifteen years, several writers have come up with positions that resemble relationism to some degree. Without attempting to be complete here, I mention some of the most obvious examples. First, the theory of natural teleology, especially as developed by Ruth Millikan (1984; 1993), is relational in stressing the aspect of utility in accounting for mental representations. Enough has been said about this theory in the previous chapter. Secondly, also the ‘experiential realism’ proposed by George Lakoff
and Mark Johnson bears a distinct resemblance to relationism (Lakoff and Johnson 1980; Lakoff 1987; Johnson 1987). Yet, it is defined in primary opposition to traditional forms of externalism (or ‘objectivism’, as they call it), which tends to make their position partially blind to the pitfalls of internalism. Similarly, the ‘ecological realism’ proposed by J.J. Gibson (1979) is clearly a form of relationism. Yet, being defined in opposition to internalism (more particularly, in opposition to classical computationalism), it is in danger of collapsing into a form of externalism itself. I come back to Gibson’s theory in the next section.

Fourthly, a number of ‘dissident’ projects in artificial intelligence, developed at MIT in the late 80s and early 90s, and called ‘interactionist programming’ or ‘Heideggerian AI’, contain essential elements of relationism. Inspired by Heidegger’s critique of the use of symbolic models of the world, they attempt to turn Heidegger’s account of skillful coping and purposive action into a research program for AI (see, for example, Chapman 1991; Winograd and Flores 1986; Dreyfus 1992, xxxff). To give an impression of this type of work, consider Chapman’s introduction of the explicitly relationist notion of deictic representation as an alternative to the traditional ‘objectivist’ approach:

“The sorts of representations we are used to are objective: they represent the world without reference to the representing agent. Deictic representations represent things in terms of their relationship with the agent. The units of deictic representation are entities, which are things in a particular relationship to the agent, and relational aspects of these entities. For example, the-cup-I-am-drinking-from is the name of an entity, and the-cup-I-am-drinking-from-is-almost-empty is the name of an aspect of it. The-cup-I-am-drinking-from is defined in terms of an agent and the time the aspect is used. The same representation refers to different cups depending on whose representation it is and when it is used. It is defined functionally, in terms of the agent’s purpose: drinking” (Chapman, 1991, 20).

Finally, also Hilary Putnam’s notion of ‘internal realism’ incorporates elements of relationism. I conclude this section by comparing my position to that of Putnam; in addition I want to make some final remarks about Kant, with whom Putnam explicitly associates himself.

In a series of well-known publications, Hilary Putnam has argued for a position called ‘internalism’ or ‘internal realism’, which he vigorously defends against all forms of ‘externalism’ or ‘metaphysical realism’ (see, for example, Putnam 1978, 123ff; 1981, 49ff; 1983, 205ff). To avoid confusion, I
refer to Putnam’s position as ‘internal realism’, and to the doctrine captured by definition (1) as ‘internalism’; similarly, ‘externalism’ is the doctrine captured by definition (2), and ‘metaphysical realism’ is the doctrine opposed by Putnam.

By ‘metaphysical realism’ Putnam means the doctrine that the world consists of self-identifying objects, that is, objects the nature of which can be specified mind-independently. Knowledge, on this view, consists in the mind’s correctly representing the intrinsic nature of the world. As an example of this position, Putnam mentions Aristotle’s theory of similitude, as well as elements of seventeenth-century theories. What is wrong with metaphysical realism is the fact that it rests on a God’s eye view of reality, which is impossible. Obviously, metaphysical realism and externalism are largely identical positions.

Putnam’s alternative is ‘internal realism’, negatively defined as the rejection of a God’s eye point of view. Positively speaking, it is the view that signs, including mental representations,

“do not intrinsically correspond to objects, independently of how those signs are employed and by whom. But a sign that is actually employed in a particular community of users can correspond to particular objects within the conceptual scheme of those users. ‘Objects’ do not exist independently of conceptual schemes. We cut up the world into objects when we introduce one or another scheme of description. Since the objects and the signs are alike internal to the scheme of description, it is possible to say what matches what” (Putnam 1981, 52; italics in original).

While agreeing with everything Putnam says in this passage, I am less attracted by the internalist overtones in his philosophy, in particular his emphasis on conceptual schemes, his rejection of naturalized epistemology (Putnam 1983, 229ff), and his association with Kantian idealism. Internal realism strikes me as a mixture of relationism and internalism. Perhaps Putnam would be happy to endorse pure relationism, and perhaps the internalist bias is only an artefact of his manner of presentation. However the case may be, I think that the following is a fair summary of the situation.

On the one hand, Putnam stresses the fact that ‘object’ and ‘representation’ essentially belong together. One does not make sense without the other: there are no ‘self-identifying’ objects without ideas to identify them, just as there are no ‘self-grasping’ ideas without objects to be grasped. This clearly
marks Putnam as a relationist. On the other hand, however, his strong emphasis on conceptual schemes, and even on transcendental reason, suggests that concepts somehow have primacy over the world conceived by them. If Aristotle took a God’s eye view of reality, Putnam seems to be taking a God’s eye view of the mind: our conceptual framework is ‘given’ to us in a sense in which reality-as-such is not given. This perspective is typical of internalism. The difference between Putnam’s relationism and his internalism is subtle but real. It is the difference between denying the point of asking what is behind our conceptual schemes (relationism), and claiming that we can never get behind the veil of concepts, or that we are locked away in our Cartesian closets (internalism). In my opinion, we should abandon Putnam’s primacy of conceptual schemes. Mind has no primacy over reality; world and mind are coeval. Similar remarks apply to Kant’s transcendental idealism, which Putnam claims to be very close to internal realism. In the previous section I pictured Kant as an internalist par excellence. I was simply following the traditional interpretation of Kant there, placing him in the tradition of radicalized Cartesianism and empiricism. There is much to be said for this interpretation. Arguably, Kant did not free himself sufficiently from the idea that conscious contents are the primarily given. In this respect, the Kantian subject is just the Cartesian mind stripped of its metaphysics. “The transcendental idealist (…) may be dualist; that is, he may admit the existence of matter without going outside his mere self-consciousness, or assuming anything more that the certainty of his representations, that is, the cogito ergo sum” (Kant 1781, A 370). Or as Richard Rorty equally densely put it, “Kant put philosophy ‘on the secure path of a science’ by putting outer space inside inner space (the space of the constituting activity of the transcendental ego) and then claiming Cartesian certainty about the inner for the laws of what had been previously considered to be outer” (Rorty 1979, 137). On the other hand, however, some lines of argument in the Critique seem to indicate that Kant was aware that neither object nor subject are ‘given’, and that it is impossible to reduce knowledge to either of its relata. This is particularly clear in his remarks on the pure categories of thought, that is to
say, the categories of the understanding stripped of their specific conditions of application.

“Now when this condition (of application to possible experience, JS) has been omitted from the pure category, it can contain nothing but the logical function for bringing the manifold under a concept. By means of this function or form of the concept, thus taken by itself, we cannot in any way know and distinguish what object comes under it, since we have abstracted from the sensible condition through which alone objects can come under it. Consequently, the categories require, in addition to the pure concept of understanding, determinations of their application to sensibility in general (schemata)” (Kant 1781, A 245; notice, incidentally, that this passage was omitted in the second edition of the Critique).

In passages like these, Kant came close to taking a relationist view of mental content. Considered in themselves, the categories of thought are on a par with things in themselves: they have no intrinsic content, containing “nothing but the logical function for bringing the manifold under a concept”. The concept-as-such, like the object-as-such, is purely ineffable: one needs the other to acquire definite meaning and form. If the relational aspect is omitted, literally no sense can be made of object and concept. In point of fact, when the twelve categories are viewed thus abstractly, they reduce to a single ‘super-category’, the so-called ‘transcendental unity of the apperception’ (op. cit., B 136ff; A 103ff). By the same token, all possible objects of knowledge reduce to a single ‘super-object’, the Ding an sich. The unity of apperception is only the abstract idea of a subject of knowledge stripped of all knowledge of its object. To turn this super-concept into a workable instrument of cognition, we need to bring in the specific conditions under which it can be applied to objects. This is where Kant applies his doctrine of schemata, mentioned earlier in this chapter. Schemata tell the understanding how the super-concept should be applied to possible experience, and thus to objects of knowledge.8

Notwithstanding the fact that Kant, who expressly rejected all traditional forms of idealism and realism, as well as the Präformationssystem of Cartesian and Leibnizian philosophy, incorporated elements of relationism in his philosophy, his general outlook was doubtless that of an internalist. Thus, his doctrine of schemata was couched in purely internalist terms, relating pure concepts to the stimuli received by sensibility, but not to something beyond.
Working within the internalist framework, Kant was ultimately unable to avail himself of the full range of conceptual resources open to relationism.

3. The bone of content

How does the relational approach outlined here bear on the theory of content? As we have seen in the previous two chapters, internalism and externalism are each in their own way committed to the claim that content is determined *intrinsically*, either by internal factors ‘given’ to the subject, or by external factors remotely controlling it. Relationism is basically the rejection of all intrinsic content. It drives home the point that there are no ‘givens’ to which to reduce the contents of knowledge. The world is not given to an external God, nor the mind to an internal Putnam.

By rejecting intrinsic content, relationism escapes from Sellars’ criticism of reductionism in naturalized theories of knowledge. In more stately terms, externalism is concerned with pure *truth*, internalism with pure *meaning*. The first wants to reduce meaning to truth, the other truth to meaning. In reality, however, there are no such things as ‘pure meaning’ and ‘pure truth’, nor can any attempt to reduce one to the other succeed. Relationism makes no attempt to reduce knowledge to something else. At the same time, however, it is also a fully naturalistic account that allows for the *reality* of knowledge. In this respect, the following remark by John Searle fits relationism surprisingly well:

“The opposite of my view is stated very succinctly by Fodor: ‘If aboutness is real, it must really be something else’ (1987, p. 97). On the contrary, aboutness (i.e., intentionality) is real, and it is not something else” (Searle 1992, 51).

An analogy may be helpful here. Consider a counterfeit twenty-dollar bill that is physically indistinguishable from the real thing. The bill is obviously worthless, for it has the wrong causal origin. Generally speaking, the value of money cannot be reduced to any of its intrinsic properties, such as its bearing the figure ‘20’. Similarly, the content of mental representations cannot be reduced to internal properties, such as features of their neural make-up. So far, the analogy rehearses a familiar argument against internalism and in favor of externalism. Yet, it may now be claimed that the value of money is still reducible to *something*, namely, to the ‘real’ or ‘intrinsic’ value of what you
can *buy* for it. In folk economics the real value of money is often explained in terms of the gold reserve at national banks. By analogy, the content of mental representations would still be reducible to something, namely, to what representations *really* go proxy for. In reality, however, there are no such things as ‘intrinsic’ or ‘real’ values, neither in economics nor in cognition. In economic reality, the value of *all* assets is a function of the laws of supply and demand, epitomized by Adam Smith as an ‘invisible hand’. What something is *worth* depends on how it can be *used* in given circumstances. Similarly, what a mental symbol *means* depends on how it can be used. This utility is determined neither by neural make-up nor by environment alone, but by the ‘invisible hand’ of cognition.

*Hybrid solutions will not do*

It is tempting to think of the ‘invisible hand’ as yet another reduction base, simply the sum of internalist and externalist factors. This is emphatically *not* the intention of relationism, however. In terms of philosophical vector space, no satisfactory solution of content determination can be found in the plane of internalism and externalism. Relationism is put forward not as a hybrid, but as a solution *orthogonal* to its predecessors.

To drive this point home, consider the difference between relationism and a typically hybrid solution. Recall from chapter six the two-tiered account of

![Diagram](image)

*Figure 8.5: Hybrid explanations of mental content*

A nucleus of ‘narrow’ or ‘computational’ content is surrounded by a cloud of additional conditions which serve to (re-)establish the link with environmental truth conditions (*left*). The diagram on the right shows the architectural companion to NC/AC semantics. For more details, see chapter six, figure 6.6.
mental content that is now popular among internalists. On this account, representational content consists of a nucleus of ‘narrow’ or computational content, surrounded by a cloud of additional constraints or ‘anchoring conditions’ (see figure 8.5). Nuclear content is what is processed computationally; it defines the taxonomy of explanatory kinds used in cognitive science. The semantic evaluation of nuclear content is supplied by externalist constraints outside the scope of cognitive science.

The hybrid account is flawed in two ways. First, its claim that the proper object of cognitive science is nuclear content is self-contradictory. Cognitive science is committed to mental content, yet nuclear content just is not content: it is what remains of a representation when all content has been bracketed. On the ‘nuclear’ approach, representations can never be studied from a genuinely cognitive point of view. Secondly, the two-tiered account tries to make up for this by supplying an account of cognitive relevance additional to the business of cognitive science. The underlying idea is something like this. Cognitive science studies the internal processing of ‘pre-semantic’ content; like the subject, it has no access to the world beyond. If we want to know what the subject is doing from an external point of view, we may zoom out to get the whole picture. We, the researchers, but not the subject, may then see how anchoring conditions fix content. This solution will not do, however. It grafts Aristotelian elements onto the Kantian component without really integrating them. The two components remain incongruously juxtaposed, instead of being conceived as essentially coherent. The relational approach, by contrast, suggests that we take the ‘anchoring conditions’ into cognitive science itself. Neither ‘internal meaning’ nor ‘external truth’ make sense when considered in isolation.

Balancing aspects of knowledge

It is not always easy to strike a clear balance between the ‘internal’ and ‘external’ aspects of knowledge. As a final demonstration of the problems involved, let me take a closer look at the theory of visual perception proposed by J.J. Gibson (1979), and at some of the criticism this theory has drawn from the internalist Establishment.

According to Gibson, the proper way to describe perception is as the ‘direct pickup’ of ‘invariant properties’ of the environment. In the case of visual perception, for example, light reverberating in the air bounces off the stable and moving surfaces in the environment and is structured by them. As a result of this, there is an ambient optic array at every point of observation, which
contains information about both the persisting layout of the environment and
the events of change that take place in it. At a moving point of observation the
structure of the changing ambient array can be described in terms of *transformations* and *invariants*. For example, when an observer is moving in a straight
line, surfaces come into view and disappear from sight as they are progressively disclosed and occluded by nearer surfaces. This information is specified
by invariants across the transformations in the optic array. To perceive the 3-D layout is simply to ‘pick up’ the invariants.

The pickup theory readily accounts for the perception of surfaces, textures, edges, objects and layouts. It is not restricted to such simple properties, however. According to Gibson, also *higher-level* properties of the environment, called ‘affordances’, are directly picked up. Affordances are properties of objects in the environment which somehow concern the goals and utilities of the subject. For example, being edible is an affordance of certain objects, as is being a tool, an obstacle, a shelter, or a predator. Similarly, objects are said to possess climb-on-ability, fall-off-ability, or get-underneath-ability (Gibson 1979, 127ff). Perception of these properties is the direct pickup of *compound* invariants.

“The central question for the theory of affordances is not whether they exist and are real but whether information is available in ambient light for perceiving them. The skeptic may now be convinced that there is information in light for some properties of a surface but not for such a property as being good to eat. (…) The skeptic understands the stimulus variables that specify the dimensions of visual perception: he knows from psychophysics that brightness corresponds to intensity and color to wavelength of light. He may concede the invariants of structured stimulation that specify surfaces and how they are laid out and what they are made of. But he may boggle at invariant combinations of invariants that specify affordances of the environment for the observer. (…) Nevertheless, a unique combination of invariants, a compound invariant, is just another invariant” (*op. cit.*, 140-141).

Finally, it is important to emphasize the *directness* of the pickup. In Gibson’s account of perception, all appeal to mediating psychological processes is eschewed: no memory, no inference, nor any other internal processing capacities are invoked in explaining perception. The pickup is emphatically *not* a matter of receiving stimuli that must be further processed. For ex-
ample, the distal layout of a scene is not *reconstructed* from the stimuli, but directly picked up in the form of invariants. Similarly, affordances such as sit-upon-ability are not *inferred* from the stimuli, but simply picked up as the ‘ready made’ properties they are.

**Against direct perception**

At first sight, Gibson’s account strongly reminds of the Aristotelian approach described earlier. Invariants are much like Aristotelian forms inherent in distal objects. Moreover, they need not in any way be processed by the subject. In terms of content determination, this seems to be just the remote control version of externalism: *what* the subject perceives is determined directly and entirely by the distal properties that are picked up from the environment. This is typically the way critics have understood Gibson’s theory. Let me briefly rehearse three lines of criticism here: one methodological, one psychological, and one conceptual.

First, it has been argued that direct perception poses a problem in scientific methodology (see, for example, Fodor and Pylyshyn 1980; Koenderink 1980; Ullman 1980). What Gibson proposes, so the objection goes, is in effect to refrain from analyzing the process of perception. He wants cognitive science to halt at the level of folk psychology. This is bad methodology, for without analysis, no explanation.

“But it is the object of science to push back the level of analysis as far as possible. This can only be done at the cost of the introduction of qualitatively new concepts. If I want to stop at the Mona Lisa’s smile, then (Gibson’s account, *JS*) is the theory for me. For the scientist a closer study (…) is compulsory. It makes the smile no less of an enigma” (Koenderink 1980, 391).

 Succinctly put, Gibson wants to explain the perception of a smile as a ‘smiling invariant’ being picked up. Similarly, the perception of a chair is ‘sit-upon-ability’ picked up. From a methodological point of view, this procedure looks suspiciously like the Aristotelian ploy of the *virtus dormitiva*.

A psychological objection against Gibson’s theory calls attention to the role of the stimulus. Ullman (1980) has argued that on Gibson’s account the information presented to the senses must be highly specific if it is to enable the organism to directly respond to it in the appropriate way. Perception becomes basically a lookup function, which immediately correlates a specific
input with a determinate output. Ullman points out that there are indeed low-
level cognitive phenomena that are immediate in this sense: for example, song
recognition in crickets is subserved by a fixed neural mechanism directly pair-
ing specific stimulus with determinate response. Most cognitive tasks, how-
ever, require much more than direct pairing; the stimulus is usually not spe-
cific enough to elicit the required response without the help of inferential,
memory-aided processing. (See the discussion of finite state machines in chap-
ter two, above.)

Finally, a related line of criticism has been launched by Fodor and
Pylyshyn in a long essay on Gibson’s theory (Fodor and Pylyshyn 1980; cf.
Schwartz 1994, 137ff). They agree with Gibson that there is information in the
ambient array of light. Moreover, they agree that it is this information by
means of which the organism learns about its environment. What they ques-
tion is Gibson’s account of how you get from the detection of features of the
light to the perception of properties of objects. Obviously, the invariants con-
tained in the light are not the same as the distal properties they are correlated
with; hence, detection of the former is not the same as perception of the latter.

“Gibson is thus faced with the problem of how, if not by inferential
mediation, the pickup of such properties of light could lead to perceptual
knowledge of properties of the environment. That is: how, if not by infer-
ence, do you get from what you pick up about the light to what you
perceive about the environmental object that the light is coming from?”
(Fodor and Pylyshyn 1980, 143).

Fodor and Pylyshyn conclude that Gibson’s account of perception is concep-
tually inadequate because it rejects in principle all internal processing.

Gibson vindicated

I believe that all of the above objections are misguided. They ignore the
relationist nature of Gibson’s proposal. It should be admitted, however, that
Gibson himself is partly to blame for this misunderstanding. He characteris-
tically exaggerated the role of the environment in cognition, and downplays
the role of the stimulus and of internal processing. I take this to be due more
to his polemical style than to the bone of his doctrine. Gibson framed his
theory in primary opposition to internalist psychology, and was particularly
keen on showing what is wrong in internalism. With the subtleties of balancing
the internal and external factors in cognition he was less concerned. I think
that Gibson’s theory can be significantly improved by explicitly discounting this balance.

Let me start with the objection raised by Fodor and Pylyshyn. From the difference between light and layout, Fodor and Pylyshyn infer that the detection of the light must be psychologically distinct from the perception of the layout. At one place in their argument this distinction appears as follows:

“the important fact is the agreement that the subject’s epistemic relation to the structure of the light is different from his epistemic relation to the layout of the environment, and that the former relation is causally dependent upon the latter” (1981, 165).

In other words, the light is assumed to be ‘given’ to the observer in a sense in which the illuminated objects are not ‘given’. If this premise is granted, the objection follows. But the premise is false: it is precisely with regard to the epistemic relations toward light and layout that Gibson refuses to make a distinction. As he puts it, a redefinition of perception is needed:

“Perceiving is an achievement of the individual, not an appearance in the theater of his consciousness. It is a keeping-in-touch with the world, an experiencing of things rather than a having of experiences. It involves awareness-of instead of just awareness. It may be awareness of something in the environment or something in the observer or both at once, but there is no content of awareness independent of that of which one is aware. This is close to the act psychology of the nineteenth century except that perception is not a mental act. Neither is it a bodily act. Perceiving is a psychosomatic act, not of the mind or of the body but of a living observer” (Gibson 1979, 239-240).

Perception is a relation between a living organism and the environment. The alleged relation to the light has no separate epistemic role additional to this relation. To be sure, the relation to the environment is realized by means of various causal processes, extending from distal object, via stimulus and proximal response, to central representation. But none of these is epistemically given, such that any of the other must be inferred.

Relationism presses the need for a balanced account of knowledge. It focuses on the overall relation rather than on any of its individual parts; none of the parts is cognitively significant unless viewed from that perspective. Thus,
detection of the light is certainly a real process, and it is part of perception. Yet, cognitively speaking, it is only an abstract phase of perception. Put in terms of the dual aspect requirement, relationism urges us to study the real parts of knowledge from an epistemic point of view.

Gibson was rather infelicitous in downplaying the psychological reality of knowledge, and of neurocomputational processing in particular. There most certainly is internal processing, but it is not conducted from something like a prisoner’s perspective. It is at this point that Ullman’s objection goes wrong. Taking Gibson’s reluctance to analyze the internal side of perception as a blunt denial of its existence, Ullman concludes that ‘direct pickup’ can only mean linear correlation between stimulus and response. Heeding the above remarks, however, a much wider range of computational resources becomes available. As discussed in chapter four, the neural nets that may arguably account for cognition and perception are much more than mere ‘lookup tables’ or ‘pattern mappers’. In particular, they involve processes of nonlinear association, and nets dynamically resonating to changing features of the environment. According to relationism, these internal processes are certainly not unreal; the point is simply that they must be studied in terms of their contribution to the overall relation between organism and environment.

This brings me to the final objection against Gibson’s theory, the charge of methodological inanity. If Gibson’s theory should indeed be seen as a fully relationist account, as I have suggested, then we should also acknowledge the transcendental nature of his proposal. A transcendental theory lays out the framework for explaining cognition without giving the explanation itself. This tallies with the account of folk psychology given in chapter three. There I argued that folk psychology gives the explananda of cognition, but not the explanation. Bound by folk psychology’s descriptive taxonomy, it is the task of cognitive science to seek for the explanations, and to push back the level of analysis as far as possible. Relationism is the perfect philosophy for this approach: a framework for the balanced integration of findings from sciences on both sides of the skin, in a truly naturalistic spirit.

4. Back to the future

*Neural epistemics and mental content*

It is time to sum up and look ahead. My leitmotif in this study has been the possibility of a naturalized epistemology. Along the way I have canvassed several aspects of cognition and mental content, including the status of folk
psychology, the power of connectionism, and the vicissitudes of content in a computational context. Let me briefly rehearse the main lines of argument here.

First, I have argued that folk psychology is indispensable for cognitive science. In this respect eliminative materialism is wrong. It is right, however, in stressing the need to study cognition from a naturalistic point of view, as a real process occurring in real material beings rather than as a purely abstract functional entity. The dual aspect theory of the propositional attitudes, introduced in chapter three, was proposed to capture this insight.

On this basis, a case was made for the large-scale integration and conceptual interaction of top-down research, which is typically concerned with the more abstract and epistemic aspects of knowledge, and bottom-up research, which is typically concerned with the neural reality of cognitive processes. The currently most suitable candidate to fit the bill is connectionism. It has the conceptual resources to bridge the gap between the buzzing confusion of the brain and the epistemic vocabulary of folk psychology. As argued in chapters four and five, the computational model of vector dynamics is a powerful tool for building a unified theory of neural epistemics, allowing us to understand the partitioning of neural vector space from a coherently cognitive point of view.

A recurrent obstacle for cognitive science proved to be the presumption that what is processed (mental content) must be intrinsic in order to be cognitive. For neural epistemics this would probably mean that content must be woven into the fabric of partitions itself. A traditional dilemma then follows. In the terminology of chapter five, neural partitions are either ‘superweak’ (empty mathematical abstractions, void of all cognitive significance), or they are ‘superstrong’ (endowed with intrinsic contents). I have argued that this dilemma can and should be resisted. How connectionism may be able to do so was demonstrated in terms of Churchland’s theory of the strong plasticity of mind.

In the last three chapters I have tried to fathom the philosophical reasons for intrinsic content. There are two competing research programs in content determination, internalism and externalism. I have argued that intrinsic content is a central tenet of internalism, but that the arguments adduced in its behalf are seriously flawed. Although externalism holds the better cards here, it ultimately succumbs to intrinsic content, too, namely, in the idea that content is remotely controlled by the environment. I pressed the need for a more balanced account of content and cognition, such that the contributions of both the
organism and its environment can be discounted. Relationism was proposed as just such an account. Like internalism and externalism, it is a transcendental framework for further research. One particularly promising implementation proved to be the theory of natural teleology.

The true significance of relationism lies in its rejection of all intrinsic content. It enables cognitive science to finally move beyond Sellars’ Myth of the Given, and toward a naturalism without the naturalistic fallacy. The proper perspective for cognitive science is neither that of Searle’s pseudo-Chinese operating the codebook of the mind, nor that of Mother Nature operating the mind’s remote control. It is the perspective of full-blooded living beings, whose interactions with the world are part of the world itself.

Beyond divine knowledge

What should be the effect of my analysis on cognitive science? First of all, I think, a new sense of philosophical awareness. Relationism invites a global reorientation in man’s image of himself. Man is being-in-the-world, not a prisoner of the body or an outside observer. Mind is more real than we commonly care to think; it is part of the same reality grasped by it, not something alien and sublime. Cognitive scientists, in my opinion, should adjust their self-image accordingly. On the standard image, they are studying universal reason under abstract conditions; on the new image, they are dealing with the human brain under human conditions.

In addition to this change in philosophical outlook, relationism calls for a rapprochement between psychology and neuroscience. Psychology needs neuroscience for its empirical constraints on real human cognition, neuroscience needs psychology for its high-level conceptual models. Connectionist theories acting as a go-between may be expected to play a role of increasing importance.

Finally, also the rejection of internalism and externalism calls for a reconsideration of research. For the development of cognitive science it will be important to turn to the world beyond the laboratory, instead of concentrating on a narrow range of artificially controlled stimuli (cf. Neisser 1976), or on the learning feats of panempiricist toy networks (cf. Dreyfus 1992). The call for more realism reminds us that internal structure is not a quirk of the brain, nor a fortuitous product of its surroundings. Rather, the quirk and the fortuity belong together.

Among connectionists it has become fashionable to speculate on the abilities of giant neural nets, equipped with massively recurrent pathways, much
more complex than the small feed-forward nets of current studies. If brains are Hegels and Einsteins, then what would a net the size of a star be? It has been pointed out that the giant net, like a brain asleep, would continue its awesome calculations even if we shut down its input and terminate its output. Should we be impressed by such speculations? I do not think so. The giant net, beautifully buzzing in its inner repose, is the connectionist image of an old friend: the Aristotelian, self-absorbed God. Divine knowledge has too long been the philosopher’s model for the human mind. We should break with this tradition. I am not sure whether divine knowledge is real, but human knowledge certainly is, and it should be studied from that point of view.
Notes

Chapter 1

1 The tension between naturalism and anti-reductionism in epistemology is particularly clear in Quine, whose attempt to combine these incongruous elements remains ultimately incoherent. For more details, see chapter 8, note 9, below.

2 In a famous passage at the outset of the *Critique of pure reason*, Kant noted that the metaphysics of his days was going through a period of crisis.

   “Time was when metaphysics was entitled the Queen of all sciences; and if the will be taken for the deed, the pre-eminent importance of her accepted tasks gives her every right to this title of honour. Now, however, the changed fashion of the time brings her only scorn; a matron outcast and forsaken, she mourns like Hecuba: *Modo maxima rerum, tot generis natisque potens—nunc trahor exul, inops*” (Kant 1781, A viii-ix; the Latin quotation is from Ovid, *Metamorphoses*, xiii, 508-510).

   To resolve this crisis, Kant submitted, philosophy needs to critically examine the possibility of metaphysics as a scientific discipline. In fine, what it needs is a general theory of the scope and validity of scientific knowledge—an epistemology as first philosophy.

3 Readers may wonder why I am apparently untroubled by a familiar argument against naturalized epistemology, namely, the objection from vicious circularity. Thus, it may be argued that knowledge is our only *access* to reality and that, therefore, the study of knowledge is logically independent of, and prior to, all studies of reality. According to this objection, epistemology must be ‘pure’: on pains of circularity, no account of knowledge can be allowed to presuppose specific knowledge about reality. I think this inference is invalid, however. First, it hinges on a notion of ‘access’ that makes the subject of knowledge an outside observer—precisely the idea with which I take issue in this study. Moreover, the ‘access’ argument seems to assume that knowledge is somehow more ‘familiar’ or more ‘intimate’ to us than the world known by it: knowledge comes first, then comes reality. Yet, knowledge is *nothing but* familiarity with the world; it is not something additional to reality, but part of reality itself. In point of fact, the argument from circularity can just as well be reversed: if knowledge is part of reality, the study of knowledge should not be biased by any prejudices about ‘pure epistemology’. I find neither the original argument nor its reversal very convincing. The first tends to overemphasize the purely logical aspect of knowledge, while the second unduly magnifies its purely natural aspect. Both of these aspects must be attended to at the same time.
Chapter 2

1 A similar argument was used earlier by Norman Kretzmann in a paper on medieval theories of the *propositio* (1970). Kretzmann’s contribution to the field of Western ethnic stereotypes involves not a Chinese but “a cooperative Turk who knows no English and German or chemistry”.

2 In his study on *Individuals*, Sir Peter Strawson introduced a sharp distinction between a form of metaphysics that is ‘descriptive’, and one that is ‘revisionary’, as he called it (Strawson 1959). The first records the conceptual underpinnings of a given domain of discourse, whereas the aim of the second (or at least its effect) is to change our scheme of things. “Descriptive metaphysics is content to describe the actual structure of our thought about the world; revisionary metaphysics is concerned to produce a better structure” (op. cit., 9). As is well-known, Strawson characterized his own work as an essay in descriptive metaphysics, claiming Aristotle and Kant as his forerunners. I think this view is misleading. *Pace* Strawson, the idea of a purely ‘descriptive’ metaphysics is a myth: all metaphysics is to some extent revisionary.

3 Similar considerations apply to the place of behaviorism in the scheme. Is it conceptually closer to materialism, or rather to functionalism? There is considerable latitude of choice. In one respect, behaviorism is a border case of functionalism (what matters is peripheral i/o profile). In another sense, it is a species of eliminativism (mental states exist only as a figure of speech). In a third sense, it is cognate to reductive materialism (every behavioral disposition determines an intervening neural mechanism). Each of these perspectives is valid from a logical point of view, and each will yield a different categorial constitution for behaviorism, that is, a different place in a differently structured conceptual space.

4 An interesting set of alternative parameters classifies theories by the nature of the cognitive architectures involved, classical or connectionist. This possibility is mooted by Andrew Clark (1987; 1989, ch. 7).

Chapter 3

1 From the day it was conceived, eliminative materialism has met with extremely varied reactions, ranging from ridicule, unbelief and contempt to exhilaration and inspired acclaim. Only part of this reception has to do with what is claimed, I think. The reactions to eliminative materialism are significantly influenced by the style of argument used by its proponents, chief among them Churchland himself. Of course, the claims are bold, but in addition they are usually stated in a *visionary* tone and supported by arguments that border on science fiction. Churchland wants ‘to boldly go where no man has gone before’, as the Star Trek generation puts it. Consider the following passage from one of Churchland’s papers, which is certainly not atypical:

“Must the journey end here? ... The long awakening is potentially endless. The human spirit will continue its breath-taking adventure of self-reconstruction, and its perceptual and motor capacities will continue to develop as an integral part of its self-reconstruction” (Churchland 1988b, 186-187)
Churchland’s tone is adroitly ridiculed by Jerry Fodor in his famous reply to the above passage that “an endless awakening sounds like not at all that much fun, come to think of it; I, for one, am simply unable to self-reconstruct until I’ve had my morning coffee” (Fodor 1988, 198).

A word of caution may be in order here with regard to the use of the term ‘methodology’. After the philosophical detour of Lakatos (1970; 1978), ‘methodology’ is here restored to what is arguably its original meaning, viz., a critical analysis of the method that is best suited for achieving certain epistemic ends (e.g., which sources should be admitted as relevant evidence, which method of demonstration is acceptable, etc.). Lakatos called it an “all-important shift in the problem of normative philosophy of science” that heuristics (“rules for arriving at solutions”) was distinguished from methodology (“merely directions for the appraisal of solutions already there”, 1978, 103 n. 1). The result, as readers of Lakatos will recall, was a system of methodologies for the appraisal of methodologies for the appraisal of methodologies for …. I do not want to enter into this hall of mirrors here. Let me state that, in my opinion, on the one hand, the appraisal of solutions already given is a matter of metaphysical and empirical success as well as of methodology, while, on the other hand, methodology should also, and in particular, be concerned with the critical evaluation of heuristics, in the sense of methods for finding solutions. The evaluation of heuristics has a double aspect: it can act in retrospect (‘Was the solution under consideration based on a sound method?’) as well as prospectively (‘Is this method likely to come up with adequate solutions?’). Whereas Lakatos’s intentions have, on the whole, been primarily retrospective, my emphasis here will be more on the prospective aspect of methodology.

On the relevant developments in epistemology and philosophy of science, several excellent introductions are now available, including those by Patricia Churchland (1986, 239-276) and William Bechtel (1988a).

Quantification over propositional contents does not carry any ontological commitment here. The reference to ‘normal circumstances’ is included to rule out cases of contradictory desires, confusion, distraction, irrationality, etc. For more ‘laws’ of this kind see, e.g., Churchland 1979, 89ff; 1981, 68ff; 1988a, 56-66. Notice that, according to some writers, folk psychological ‘laws’ of this kind cannot be empirically defeated, for the simple reason that people who are not disappointed in case (6) either do not fill the correct antecedent or are behaving irrationally. If you fail to follow the rule, then your behavior is at fault, not the rule. I am not going into this problem here, as it will not affect my main argument. (For a discussion, see, e.g., Dennett 1981a and 1981b; Stich 1981, 1985 and 1990; Cohen 1981.)

The arguments given here are schematic reconstructions of various lines of reasoning and scattered remarks gleaned from Churchland’s books and papers, listed in the bibliography. Some of his most important papers on the subject have been collected in Churchland 1989a.


A suggestive way of explaining the observationality of folk pronouncements is in terms of Quine’s remarks on grades of theorecticity and observationality (see, e.g., Quine 1970; 1990, 2-6). As Quine put it:
“The quality of being an observation sentence does, strictly speaking, admit of degrees. Thus we may allow for one speaker’s failure to agree with other speakers in assenting to an observation sentence on some occasion; we may simply take the degree of rarity of such exceptions into consideration in assessing how observational a sentence is. In further refinement of the notion, we may allow a speaker to revoke his assent to an observation sentence, or to revoke his dissent, after additional stimulation. (...) The degree of observationality of a sentence might then be measured inversely by the average dose of stimulation needed to induce a stable verdict” (Quine 1970, 5).

8 It is sometimes claimed that the best candidate for a textbook on folk psychology would be Ryle’s famous work, *The concept of mind*. I venture to disagree with this choice, preferring the novels by C.P. Snow, such as his Lewis Eliot sequence (also known as *Strangers and Brothers*), to be given this place of honor for their brilliantly penetrating analysis of motives, scruples, anxieties, politics and intrigues.

9 Stephen Stich takes this argument against sentential parameters to be “justified only when ... directed towards theories which take cognitive states to be contentful or semantically interpreted mental sentences” (1983, 217). In other words, he takes it to be an argument for the formality condition, which will be discussed in the next chapter. On this reading, Stich’s own ‘syntactical’ theory of mind is unaffected by the argument. I disagree with this interpretation of Churchland’s position. If sentential parameters are rejected, then so are syntactical sentential parameters. One as well as the other are a matter of sentence crunching.

10 The question whether folk psychology is bound to sententialism should be sharply distinguished from that of whether theorizing in cognitive psychology is sentential. Yet, even with regard to the second question Churchland’s claims can be disputed, as Barbara Von Eckardt (1984) has pointed out. She draws attention to the work of Stephen Kosslyn, whose theories of mental imagery invoke the existence of “mental representations which he has been valiantly attempting to distinguish from discursive representations” (**op. cit.**, 85). Notice also that many philosophers, most notably Fodor, explicitly base themselves on folk psychology in support of the sentential paradigm in cognitive psychology. For a straightforward argument to this effect, see Fodor 1987, 135ff. I shall come back to this argument in chapter six. For a discussion of the strained relation between sententialism and **connectivism**, see chapter four, below.

11 Important and well-known exceptions to this rule include the top-down strategy followed by neuropsychologists in the field of syndrome analysis (for an excellent introduction to the subject, see, e.g., Luria 1974).


13 A similar argument can be used to demonstrate that folk psychology does not make any specific claims about the nature of consciousness. In particular, it does not commit us to the view that consciousness is a place in the mind/brain where all information comes together for inspection by some kind of ‘Inner Observer’. Thus, in my opinion, Daniel Dennett’s recent argument against the Myth of the Cartesian Theatre does not by itself impugn folk psychology (Dennett 1991; cf. also Patricia Churchland 1983).
Chapter 4

1 For a brief history of connectionist theory, from Pandemonium and Perceptron models to present-day neural nets, see Rumelhart and McClelland 1986a, vol. 1, 41-44; Bechtel and Abrahamsen 1991, ch. 1; Boden 1991; Churchland and Sejnowski 1992, ch. 3.


3 I thank my student Roland Peterson for a number of helpful suggestions on the material presented in this section.

4 The turn from instant associator to dynamic system was certainly an important breakthrough in the development of connectionism. In particular, it has been suggested that John Hopfield's interest in connectionist theory was crucial for its becoming a legitimate science. Thus Anderson and Rosenfeld comment:

"John Hopfield is a distinguished scientist. When he talks, people listen. Theory in his hands becomes respectable. Neural networks became instantly legitimate, whereas before, most developments in networks had been in the province of somewhat suspect psychologists and neurobiologists, or by those removed from the hot centers of scientific activity" (Anderson and Rosenfeld 1988, 457).

It should be clear from the above that the advent of new mathematical techniques for describing complex net dynamics was indeed an all-important contribution to connectionism—if only because it enabled theorists to explore new and much more complex architectures than before.

5 In a sense, classical cognitive models can thus be said to endorse the same positive heuristics as connectionist theories. For an account of plasticity in cognitivist models, especially with regard to learning processes, see Klahr 1987. Modularity is a feature of virtually all models of cognition; for classical accounts, see Fodor 1983; Pylyshyn 1984; Gopnik and Gopnik 1986; Garfield 1987. Finally, as for the use of distributed representations in classical cognitivism, it should be noticed that microfeature analysis is not the privilege of connectionism. Virtually all accounts of semantics endorse the technique in one form or another, though not uncontroversially so. Thus, microfeature analysis is a popular technique in phonetics, in structural linguistics (Lyons 1968), and in theories of memory (for an overview, see Smith and Medin 1981). Cf. also Pinker and Prince 1988b.

6 For some excellent introductions to the philosophically most salient aspects of Marr's theory of vision, see Boden 1984 and Kitcher 1988; also, Burge and Stillings in Garfield 1987, 365ff and 383ff.

7 In an ideal situation we would be able to first train up a *tabula rasa* net and then analyse its acquired cognitive structure. This would save us the trouble of a prior and laborious analysis of the relevant task domain. Instead of building innate knowledge into the net's connectivity structure, its acquired structure would teach us something about our innate (or acquired) structure. In present-day research, much energy is invested in this unduly optimistic
approach, as is testified, e.g., by the methodological suggestions of Smolensky 1987 and 1988, 8. For a discussion of some of the problems posed by ‘panempiricist’ nets, see Dreyfus 1992, xxxvii ff.

8 In his seminal book, *The society of mind*, Marvin Minsky has put this question as follows (notice that he uses the term ‘insulated’ in the sense of ‘modular’ described above).

“The advantages of distributed systems are not alternatives to the advantages of insulated systems; the two are complementary. To say that the brain may be composed of distributed systems is not the same as saying that it is a distributed system – that is, a single network in which all functions are uniformly distributed. I do not believe any brain of that sort could work, because the interactions would be uncontrollable. To be sure, we have to explain how different ideas can become connected to one another – but we must also explain what keeps our separate memories intact. (…) Some theorists have assumed that distributed systems are inherently both robust and versatile, but actually those attributes are likely to conflict. Systems with too many interactions of different types will tend to be fragile, while systems with too many interactions of similar types will be too redundant to adapt to novel situations and requirements. Finally, distributed systems tend to lack explicit, articulated representations, and this makes it difficult for any such agency to discover how any other such agency works” (Minsky 1985, 319-320).

Note that this passage establishes a link between the concept of modularity and that of distribution of representation. For a colorful and provocative application of Minsky’s theory, see also the science fiction novel by Harry Harrison, coauthored by Minsky himself, *The Turing option* (1992). Interested readers will find persuasive (though admittedly speculative) applications of a number of points made in Minsky 1985.


10 The locus classicus for a discussion of the concept of ‘distribution’ (the ‘D’ in ‘PDP’) is Hinton et al. 1986. See also Van Gelder 1991; Lloyd 1989, 102ff; Clark 1989, chs. 5 and 6; Churchland and Sejnowski 1992, 163ff.

11 Plasticity and learning capacity as exhibited by connectionist nets have considerable neural plausibility. Different kinds of long-term structural change have been distinguished at the neural level, including long-term potentiation (LTP), activity-dependent facilitation (ADF), structural connectivity modification (SCM), and central control or mnemonic learning. In 1973, Timothy Bliss and Terje Lømo first demonstrated that neurons in the hippocampus show remarkable plasticity of the kind that would be required for forms of explicit learning. For more details, see Kandel and Hawkins 1992; Churchland and Sejnowski 1992, 239ff; Changeux 1985, 205ff.

12 Vector coding effectively takes care of the ‘running-out-of-neurons’ objection against local coding by means of the following combinatorial consideration. If \( n \) binary nodes can locally code \( n \) different symbols, then they can vector code \( 2^n \) different distributed symbols. Moreover, if \( n \) nodes have \( w \) connections each, and if each connection can take \( v \) discrete values, the net’s overall weight matrix, which determines its stored vectors, can take \( v^{nw} \) different configurations.
Notice, incidentally, that these examples may arguably be taken as corrections of philosophy of science, rather than of folk psychology—another instance of folk psychology’s profound theoretical innocuity.

Arguments along this line are spawned by the position of ‘subsymbolists’ (see, for example, Hofstadter 1985; Smolensky 1987 and 1988; Cummins 1989, ch. 11). The objection sketched in the text is based on suggestions by Robert Cummins (1989, ch. 11), who has now retracted these earlier claims (see Cummins 1991, 104).

Notice, incidentally, that the purported primacy of classical machines is \textit{not} forced upon us by folk psychology, as Fodor and Pylyshyn suggest. In chapter three I argued that the paucity of empirical constraints, such as posed by connectionist architecture, in attributions of propositional attitudes does not mean that these constraints are irrelevant; rather, it is a reminder that they still need to be filled in.

\textit{Chapter 5}

1. Well-known objections against physicalism, based on a consideration of the specific properties of qualia, include those of Kripke 1972; Nagel 1974; Block 1978; Searle 1980; Jackson 1982; Robinson 1982. For a colorful collection of thought experiments for and against physicalism, see Hofstadter and Dennett 1981. As the ‘mother of all arguments’ in this field, I mention Kripke’s modal argument. A rough version of it runs as follows. Qualitative states of consciousness, such as pain, are ruled by conditions of identity that are totally different from those determining the identity of physical states, such as states of the nervous system. Hence, for all neural states \(N\) and conscious states \(C\), it is perfectly conceivable for a subject to be in state \(N\) without being in state \(C\). Put in philosophical terms, if \(N\) is identical with \(C\), this identity will be contingent at best. The identity of \(N\) and \(C\) is not projectible across possible worlds, and thus it fails to meet one of the main requirements for scientific identity. Scientific identifications, such as that of water and \(H_2O\), are necessary and projectible: it is inconceivable that something is water but not \(H_2O\). Kripke’s argument inspired a surge of thought experiments, in all of which the phenomenal identity of states of consciousness is kept constant, while internal (neural, functional) or external (ecological, historical, social, linguistic) factors are allowed to vary. On the reign of philosophical surrealism released by this style of argument, see also chapter two, above, and this chapter, note 5.

2. In his recent book on the \textit{Pursuit of truth}, Quine remarks that “some sentences (…) are directly and firmly associated with our stimulations. (…) I call them observation sentences. (…) Observation sentences are the link between language, scientific or not, and the real world that language is all about” (Quine 1990, 2ff). Mark that the weaker version of network semantics is cognate to the position adopted in chapter three with regard to the status of folk theory. See also chapter three, note 8.

3. Psychological empiricism, of the kind advocated by Locke, was identified by Kant as “the belief that the senses not only supply impressions but also combine them so as to generate images of objects (man glaubte, die Sinne lieferten uns nicht allein Eindrücke, sondern setzten solche auch so gar zusammen). For that purpose something more than the mere receptivity of impressions is undoubtedly required, namely, a function for the synthesis of them” (Kant 1781, A 120, note). Notice that it is irrelevant to my general line of argument whether or not
the position outlined here is historically correct as a rendition of Locke's theory. For a
defense of Locke as a 'direct realist', see Mandelbaum 1964, and, more recently, Yolton 1987.
For a related discussion of 'radical realism', see Pols 1992.

4 Churchland’s choice of words is very close to Sellars’s original phrase, that “in characterizing
an episode or state as that of knowing, we are not giving an empirical description of
that episode or state; we are placing it in the logical space of reasons, of justifying and being
able to justify what one says” (Sellars 1963, 169, italics mine).

5 It is important to note that examples such as these are, in reality, nothing but rhetorical
devices which may help us to imagine what it would be like to have infrared eyes, ceteris
paribus. The cetera contain the venom: the more exotic the assumptions, the more feeble our
intuitions tend to be. Suppose our normal retinala are replaced by a tissue that is sensitive
to radiation in the far infrared, while the rest of our visual system remains intact — how
would the world then look to us? The question becomes considerably more difficult to
answer if we contemplate the effects of additional surgical changes, including a new connection
between our new eyes and the old neural vector space for tactile perception of
temperature. Would we still be seeing the same things, or would we rather be feeling them?
Personally, I must frankly admit that my intuitions fail me in cases like these. On the reign
of surrealism in modern philosophy of mind, see also Wilkes 1988, 1-48 ('Personal identity
without thought experiments'), and note 5, above.

6 The analysis given here is allied to Churchland’s defense of functionalism against qualia-
based objections (1988a, 38-42). Churchland suggests the following solution: “identify the
qualitative nature of your sensations-of-red with that physical feature (of the brain states
that instantiates it) to which your mechanisms of introspective discrimination are in fact
responding when you judge that you have a sensation-of-red. If materialism is true, then
there must be some internal physical feature or other to which your discrimination of sen-
sations-of red is keyed: that is the quale of your sensations of red” (op. cit., 40).

7 In A natural history of the sense, the American naturalist and poet Diane Ackerman sharply
observes:

“When light hits a red car on the streetcorner, only the red rays are reflected into our
eyes, and we say ‘red’. The other rays are absorbed by the car’s paintjob. When light
hits a blue mailbox, the blue is reflected, and we say ‘blue’. The color we see is always
the one being reflected, the one that doesn’t stay put and is absorbed. We see the re-
jected color, and say ‘an apple is red’. But in truth an apple is everything but red”
(Ackerman 1990, 252).

For a strongly reductionist account of sensory qualia, integrating findings from psycho-
physics, psychometrics and sensory neurophysiology, see also Austen Clark 1993. Clark’s
discussion covers chromatic perception as well as other modalities, including taste, smell
and spatial qualia. The notion of triplet coding for color sensations, as functions of (blue,
green, red), goes back to Edwin Land’s so-called ‘retinex’ theory of color vision; for a first
introduction, see, e.g., Land’s well-known paper in Scientific American (Land 1977). Edwin
Land was not a biologist, but an engineer who invented the instant camera and Polaroid
filters. More importantly, he was also a keen observer of how we observe. Land’s theory is
based on the idea that colors are judged less by the relative stimulation of neighboring ‘red’,
‘green’, and ‘blue’ cone receptors, than by the ratio of local and global intensities for red,
green and blue. The global, across-image lightness in each of the reflectancy bands can ex-
perimentally be shown to affect the quality of the color that is perceived (Gregory 1987,
q.v.). Together, the three values for global lightness provide the coordinates of a three-dimensional space, such as shown in figure 5.4. Whereas a color space based on the absolute absorption values in the three classes of receptors will predict only whether or not two physical stimuli will match, a space based on the three lightnesses of retinex theory will predict how colors actually look to us. For anti-reductionist criticism of the retinex theory, see Campbell 1982.

8 Alternatively, from a slightly different perspective, it is equally correct to say that qualia do not wear labels at all, but that they are the labels themselves.

9 Jackson is concerned with a different question, namely, what kind of sensation Fred is having when he sees red₁ or red₂. “What is the new color or colors like? We would dearly like to know but do not; and it seems that no amount of physical information about Fred’s brain and optical system tells us. (…) It follows that Physicalism leaves something out” (Jackson 1982 (1990, 470-471)).

10 An attempt to combine the best of both worlds, that of theory and that of observation, is made by Fodor in his suggestion that perceptual systems are organized as modules, in the technical sense explained in chapter two (Fodor 1983; Fodor 1988). According to Fodor, the ‘perceptual modules’ that are responsible for observation are rigid: their built-in theories operate in fixed ways that cannot be altered by any background knowledge supplied by the exploiting theories. This feature of ‘informational impenetrability’ would make observation theory laden but not plastic. For further discussion, see Churchland 1988b; Meijering 1989b; Ross 1990.

11 This seems to me to be one important way to read Paul Feyerabend’s dense paper on ‘science without experience’: “sensations can be eliminated from the process of understanding (…) (though they may of course continue to accompany it, just as a headache may accompany deep thought” (Feyerabend 1969 (1981, 134)).

12 Kant considered it to be a “scandal to philosophy and to human reason in general (ein Skandal der Philosophie und allgemeinen Menschenvernunft) that the existence of things outside us (from which we derive the whole material knowledge, even for our inner sense) must be accepted merely on faith, and that if anyone thinks good to doubt their existence, we are unable to counter his doubts by any satisfactory proof” (Kant 1781, B xxxix, note). His refutation of psychological idealism was intended as a remedy for this outrage.

13 Cf. Aristotle, Metaphysics XII, 1074b34-35. Thomas Aquinas, Summa Theologiae I, q. 87, remarked that the intellect has knowledge of itself only by reflection on its knowledge of other things.

Chapter 6


Due to the scholastic nature of the discussion, the theory of content often looks like a game of intellectual pingpong. The effect is due largely to the subtlety and rhetorical talent of the main contributors, first among them Jerry Fodor. After ten years of apparently fruitless discussion, signs of wear and despair are beginning to show. A characteristic example is the tone of Fodor’s reply to criticism in Loewer and Rey (1991a, 255ff).

“I’m at a bit of a loss how to proceed with the Baker paper (Baker 1991, JS). It raises a lot of detailed worries about the theory of content in ‘A theory of content’ and, though I don’t think her objections actually cut ice, this stuff does get a little complicated; points need to be made that only a devotee of informational semantics could conceivably care about. Other readers are likely to go berserk and do themselves a harm, and I do not wish to have their blood on my hands” (Fodor 1991, 258).

2 My criticism here can be seen as a necessary complement to Dennett’s attack on the ‘Cartesian Theatre’ (Dennett 1991; Dennett and Kinsbourne 1992). Dennett attempts to show that much modern theorizing about the mind is still wed to a Cartesian notion of consciousness as the arena in which all mental content gathers for inspection by the ego. If Dennett is arguing that there is no such mental stage, or no mental audience, I want to cast doubt on our standard conception of the actors.

3 As is well-known, the separation of processing structure and memory (including input cards and master cards) was one of the major breakthroughs in computer science, due to John Von Neumann. The separation poses specific constraints on computational architecture. For a discussion of the distinction between processing structure program and data in a psychological context, see, e.g., Pylyshyn 1978, 1980 and 1984. David Hillis, one of the developers of the Connection Machine, points out that the Von Neumann design was motivated largely by considerations of technology and economy:

“When von Neumann and his colleagues were designing the first computers, their processors were made of relatively fast and expensive switching components, such as vacuum tubes, whereas the memories were made of relatively slow and inexpensive components, such as delay lines or storage tubes. The result was a two-part design that kept the expensive vacuum tubes as busy as possible. We call this two-part design, with memory on one side and processing on the other, the von Neumann architecture, and it is the way that almost all computers are built today. This basic design has been so successful that most computer designers have kept it even though the technological reason for the memory/processor split no longer is justified” (Hillis 1985, 4).

Nowadays, processor and memory in standard computers are made of the same material, namely, silicon chips. The relative inefficiency of memory as compared to processor inherent in the Von Neumann design (the so-called ‘Von Neumann bottleneck’) was one argument for designing a new type of computer, the Connection Machine.

4 The differences between these types of properties are not absolute. Consider a conventional symbol such as a written letter ‘S’. Although both its shape and size are, strictly speaking, formal properties in the sense of definition (2), we readily think of shape as somehow being more syntactic, that is, being closer to the way the symbol will be manipulated computationally. Similarly, in the case of punch cards, both their weight and their punch pattern are
formal properties. Still, we tend to think of punches as being relatively close to syntactic properties, whereas weight is more readily relegated to the realm of hardware. (Yet, some machines may utilize precisely the weight of cards for encoding information, ignoring whatever holes they may contain.) Again, consider the shape of an individual symbol such as ‘BANK’, as opposed to the way it may be combined with other symbols to form sentences according to the rules of grammar. Although, strictly speaking, both are syntactic properties, involving the rules for combining letters to form words, or for combining words to form sentences, the shape of a word intuitively strikes us as being closer to the formal level than to the semantic level, while the opposite is true of the ways in which the ambiguous term ‘bank’ can be used in various contexts.

5 Pace Putnam, philosophers such as Husserl would probably opt for an interpretation on which the ‘water’ thoughts of Mary and Mary∗ have identical contents in virtue of their identical phenomenology. Moreover, there is some evidence for this option from recent theories in psycholinguistics. Psychologists such as Eleanor Rosch have argued for a view on which a single concept may be realized by multiple mental representations, distinguished by the specific context of purposes for which they may most profitably be applied (see, e.g., Rosch 1978; Smith and Medin 1981; Roth and Frisby 1986; Smith 1990). The concept ‘gold’, for example, may be represented as either (a) yellow, shining metal, (b) material that is malleable, valuable, and a conductor for electricity, (c) element with atomic number 93, and so on for other specific descriptions. According to Rosch, which of these specific ‘sub-representations’ a subject will use depends on the circumstances. From this point of view, it may be argued that Twins share a common, folk-theoretical ‘sub-concept’ of water, which, upon the discovery of chemical structure, will come to be associated with different scientific ‘sub-concepts’. Notice, however, that the ‘sub-representation’ theory is essentially different from ‘narrow content’ analyses like Fodor’s. For a discussion of the proper interpretation of Rosch’s data, see also Lakoff 1987.

6 To give an impression of the state of confusion in the pertinent literature, let me sample some of the best-known terminology here. Externalist theories have also been called naturalistic psychology (Fodor 1980), representational theory of mind (Stich 1983), non-individualism (Burge 1986), naturalism (Stillings 1987; Sterelny 1990), external relational theory (Loewer and Rey 1991b), wide functionalism (Kitcher 1985), non-supervenient psychology (Kim 1982; cf. Davidson 1970), or just folk psychology (Stich 1983); with regard to their objects, they have been identified as the theory of transparent taxonomy (Fodor 1980), informational content (Dretske 1981), calibrational content (Churchland and Churchland 1983), propositional attitudes, broad or truth-conditional content (Fodor 1987), wide beliefs, wide content, or psychological states in the wide sense (Putnam 1975). Internalist theories, by contrast, have also been called methodological solipsism (Fodor 1980; Kitcher 1985), syntactic or autonomous psychology (Stich 1983), individualism (Burge 1986; Sterelny 1990), narrow functionalism (Kitcher 1985), or supervenient psychology (Kim, Davidson); with regard to their objects, they have been identified as the theory of opaque taxonomy (Fodor 1980), intentional content (Dretske 1981), translational content (Churchland and Churchland 1983), narrow content (Fodor 1985b; 1987), functional/conceptual role (Block 1986), or psychological states in the narrow sense (Putnam 1975), satisfying the formality condition (Fodor 1980). Von Eckardt (1993, ch. 7) distinguishes three constraints on content determination that are at issue here, namely, naturalism, internalism, and methodological individualism.

7 The origin of the term ‘supervenience’ lies in moral philosophy (G.E. Moore, Philosophical studies, 1922). In the following passage from R.M. Hare (The language of morals, 1952, p. 145),
supervenience is explained in terms of a replacement argument, which in the literature today is still its most characteristic form.

“First, let us take that characteristic of ‘good’ which has been called its supervenience. Suppose that we say ‘St. Francis was a good man’. It is logically impossible to say this and to maintain at the same time that there might have been another man placed in exactly the same circumstances as St. Francis, and who behaved in exactly the same way, but who differed from St. Francis in this respect only, that he was not a good man.”

Notice that supervenience, in this passage, is emphatically not concerned with the bodily substrate of St. Francis (the man as such), but rather with his circumstances and behavior.

8 For a brave put puzzling attempt to hold on to the Cartesian assumption, see Searle 1992.

9 As a matter of fact, Putnam’s original Twin-Earth argument against methodological solipsism specifically addressed the ‘California’ version of conceptual role semantics; see Putnam 1975b, 262ff. For meaning holism and ‘confirmational’ or epistemic holism, see Fodor 1987, ch. 3; Fodor and Lepore 1992. The latter work reviews versions of holism by Quine, Davidson, Lewis, Dennett, and Block. Cognate features of belief fixation, most notably the ‘Quinean’ and ‘isotropic’ character of confirmation in science, are discussed by Fodor 1983, 104ff (see also Garfield 1987, and chapter two, above). An issue of related interest is the role of rationality in the attribution of beliefs (see chapter three above, note 4). The classical introduction to referential opacity in epistemic contexts is Chisholm 1957. Arguments for internalism based on opacity include Fodor 1980; Stich 1983, ch. 6. For a discussion of these arguments, see, for example, Dennett 1982 (1987, 174ff); Garfield 1988, 57ff. Arguments from conceptual change are often raised in combination with holism; see, for example, Block 1986; Fodor 1987, ch. 3; Fodor and Lepore 1992; for a non-internalist discussion of conceptual change, see Churchland 1979, ch. 3.

Chapter 7

1 For a general introduction to externalism in the philosophy of mind, see Putnam 1975b; Burge 1979, 1986. For a defense of externalism from a psychological point of view, see Gibson 1979; Neisser 1976. In section two, below, I distinguish three main types of externalist or so-called ‘causal’ accounts of mental content, namely, actualist theories, counterfactual theories, and teleological theories. The two most influential ‘actualist’ theories were originally introduced by Saul Kripke (1972) and Hilary Putnam (1975b). Early examples of covariational theories are Stampe 1979 and Dretske 1981. For an excellent discussion of Dretske’s position, see Possin 1984, which adds several important teleological improvements to the counterfactual theory. Critical studies of Dretske’s view include Fodor 1984, Taylor 1987, and Morris 1990. Since his 1981 book, Dretske has refined his theory, with particular regard to the problem of misrepresentation, in Dretske 1986 and 1988. In addition to his internalist views, Fodor, too, has advocated various kinds of causal theory, both correlational and teleological; see, for example, Fodor 1984, 1986, and 1990d. All of these earlier theories are now officially rejected by Fodor himself; for his severe criticism of his former self, see Fodor 1987, 1990b, and 1990c. Influential examples of teleological theories include Fodor 1984 and 1990d; Millikan 1984, 1986, 1989, 1990 and 1993; Papineau 1987 and 1993, ch. 3. Various aspects of externalist accounts are discussed by Churchland and Churchland

2 For a particularly lucid discussion of idealization in covariational theories, see Cummins 1989, ch. 4. Also noteworthy is Cummins’s account of idealization and teleology, op. cit., 78ff.

3 For some classical accounts of misrepresentation and the disjunction problem in causal theories of content, see, for example, Fodor 1984 and 1987, ch. 4; Dennett 1987, ch. 8; Dretske 1988, 64ff; Cummins 1989, ch. 5. Misrepresentation is also addressed by several of the papers collected in Loewer and Rey 1991a.

4 The reference is to Shannon and Weaver 1949. For a critical examination of the concept of information used in their theory, and its philosophical utility in Dretske’s account of knowledge, see the discussion in the Open Peer Commentary on Dretske 1983; Taylor 1987; Morris 1990.

5 Other difficulties with Dretske’s proposal include the problem of how to identify, in a non-circular and principled way, the period during which learning takes place. Also, the application of Dretske’s approach is necessarily restricted to acquired concepts; it is unclear how it can be worked into an account of the semantics of innate concepts. Moreover, Dretske’s solution entails a sharp distinction between natural and intentional representation. Finally, it is unable to account for cases in which misrepresentation is in fact functional—for example, in situations when it pays to be cautious rather than accurate (cf. Fodor 1984, 241; Taylor 1987). For a defense of Dretske’s solution, introducing several important teleological corrections, see also Possin 1984, 130ff. Dretske himself has abandoned his original position with regard to misrepresentation; see Dretske 1986 and 1988, 64ff.

6 In addition to this simplified version of his solution, Fodor also carries a ‘parade version’ (1987, 164, note 6). This more elaborate version is essentially as follows: “B-caused ‘A’ tokens are wild only if the nomic dependence of instantiations of the property of being an ‘A’ tokening upon instantiations of the property of being a B tokening is itself dependent upon the nomic dependence of the property of being an ‘A’ tokening upon instantiations of some property other than B.” The objections raised in the text hold for the parade version as well as for the more simple version.

7 Much the same conclusion can be reached if we examine Fodor’s discussion (1990c, 103ff) of an example raised by Lynne Rudder Baker (Baker 1989 and 1991). Baker invites us to consider a case that is only slightly different from the one discussed above. Imagine a set of robot-cats impinging on sensory surfaces in exactly the same way as real cats. Suppose further that a person S learns a particular concept C solely from exposure to robot-cats. Now, Baker asks, what will be the content of C? Three possibilities suggest themselves:

- (i) C bears the content \(|\text{cat}|\), and robot-caused C-tokens are wild;
- (ii) C bears the content \(|\text{robot-cat}|\), and cat-caused C-tokens are wild;
- (iii) C represents the disjunctive property of being either a cat or a robot-cat.

Somewhat surprisingly, given his reaction to the cow/horse example, Fodor’s initial response here is to endorse possibility (iii): under the particular circumstances described by Baker, the subject S has not yet learned the distinction between cats and robot-cats. Fodor then proceeds to examine the disjunction problem “a little more carefully than one usually
needs to” (op. cit., 105). If S cannot tell the difference between robots and cats, does this mean that he has a disjunctive concept \( \text{cat or robot} \)? According to Fodor, the answer is negative. “Nobody can have the concept CAT OR ROBOT unless he has the constituent concepts CAT and ROBOT; which by assumption, S didn’t” (loc. cit.). Obviously, the appeal to constituency begs the question of content. See also chapter four, above.

Teleological accounts of content include Millikan 1984, 1986, 1989, 1990 and 1993; Fodor 1984 and 1990d; Papineau 1987 and 1993; Dretske 1988. For a critical discussion of teleological accounts, see, for example, Cummins 1989, ch. 7; Fodor 1991b, 64ff; Von Eckardt 1993, 223ff. Fodor’s trail in the field of teleology is characteristically confusing—the Fodorian enigma never ceases to amaze. In *Psychosemantics* (1987), Fodor severely criticized the teleological theory he had proposed earlier in a manuscript bearing the same title, widely circulated from 1984-1985 onward. Fodor has only reluctantly allowed this manuscript to be ‘posthumously’ published (Fodor 1990d), “for he now thinks the view it defends to be hopelessly and viciously wrong” (Fodor 1990d, 312, note).

As a matter of fact, the ‘subjective’ constraint on teleology also seems to hold for ordinary measuring instruments. We tend to think, rather naively, that the calibration of measuring instruments is only a matter of establishing perfect correlations under normal conditions. That the instrument was designed specifically to fit our purposes is usually forgotten. Yet, this subjective constraint weighs heavily on the determination of the content of our measurements. For example, we usually think that a speedometer is calibrated to yield correct representations of our velocity under normal conditions. In reality, however, its readings are almost never veridical, not even under normal conditions. The meter presents its information in a format that is characteristically useful to us, but that is rarely accurate in a physical sense. Accurate representation would cause the dial to vibrate restlessly up and down, making it far too difficult to read. For purposes of ergonomics, the dial is balanced so that it floats gently over the scale. Here, too, utility is at least as important as physical correlation for the determination of representational content.

At other times, however, Fodor seems to be denying this conclusion, arguing that \( \text{ambient black dot} \)s may not be reliably equivalent to \( \text{retinal state P} \)s, or at least that, generally speaking, content ascriptions in terms of specific distal causes may not be equivalent to content ascriptions in terms of specific proximal causes. Consider, for example, the following reflection:

“If all this is true, then the frog’s fly-elicited fly snaps are asymmetrically dependent on these states of retinal excitation. So why aren’t the excitation states the intentional objects of the frog’s snaps? I don’t know what the story is with frogs, but in the general case there is no reason to suppose that the causal dependence of perceptual states on distal objects is asymmetrically dependent on the causal dependence of specific arrays of proximal stimuli on the distal objects; e.g., that there are specifiable sorts of proximal traces that a cow has to leave on pain of the cow-COW connection failing” (Fodor 1990c, 109).

I am very much puzzled by this move. As far as I can see, if dot-P equivalence is denied, the category of ambient black dots simply loses its coherence. (Even flying flagpoles, suitably projected onto the retina, belong to the category of ambient black dots. If this is not by virtue of their causing P, I can think of no other reason why we should call them black dots.)
Chapter 8

1 Cummins (1989, ch. 1) notes that the number of different solutions that have been tried with regard to the problem of content determination is surprisingly small. He distinguishes four such solutions: similarity, covariance, adaptational role, and functional or computational role. While agreeing with this choice of possibilities, I think it fails to bring out some essential points of difference and agreement. Similarity, covariance and adaptational role are typically externalist options, while functional/computational role is typically internalist. In this respect, my estimate of the number of different solutions is even more conservative than that of Cummins.

2 I give here the relevant texts on which the diagram is based. Aristotle, *De interpretatione*, I, 1-3: “Sunt ergo ea quae sunt in voce, earum quae sunt in anima passionum notae et ea quae scribuntur, earum quae sunt in voce. Et quemadmodum nec litterae omnibus eaedem, sic nec eaedem voces; quorum autem hae primorum notae, eaedem omnibus passiones animae sunt, et quorum hae similitudines, res etiam aequadum.” Boethius, *In de interpretatione*, IIa editio, I, 1: “Res enim ab intellectu concipitur, vox vero conceptiones animi intellectusque significat, ipsi vero intellectus et concipient subjectas res et significantur a vocibus. Dicendum est res et intellectus, quoniam apud omnes idem sunt, esse naturaliter constitutos, voces vero atque litteras (...) non esse naturaliter sed positione. (...) Nam cum ea quae sunt in voce res intellectusque significat, principaliter quidem intellectus, res vero quas ipsas intelligantia comprehendit, secundaria significacione per intellectum mediatement.”

3 See also chapter five, n. 3.

4 There are important relational, and more particularly teleological elements in Aristotle’s theory of perception and cognition (see, e.g., Block 1961). Similarly, Kant’s internalism incorporated elements of a distinctly relational approach, to which I presently return. Yet, speaking in terms of historical and conceptual impact, Aristotelian psychology is best seen as externalist, Kantian epistemology as internalist.

5 In point of fact, Kant distinguished three kinds of deduction with regard to our use of concepts: empirical, metaphysical, and transcendental. The empirical deduction “shows the manner in which a concept is acquired through experience and through reflection upon experience, and which therefore concerns, not its legitimacy, but only its de facto mode of origination” (Kant 1781, A 85, B 117). Kant also used the term ‘metaphysical deduction,’ albeit sparingly. He meant by it the manner in which “the a priori origin of the categories has been proved through their complete agreement with the general logical functions of thought” (op. cit., B 159), referring to the cue to the discovery of all pure concepts of the understanding (op. cit., A 66ff, B 91ff). Finally, a transcendental deduction was defined by Kant as “the explanation of the manner in which concepts can (...) relate a priori to objects” (op. cit., A 85, B 117). As opposed to a mere empirical deduction, the transcendental deduction is concerned not with the de facto origin of our concepts, but with their legitimacy. More precisely, the transcendental question is not under which empirical conditions the application of a given concept X is justified, but rather how the application of any concept can be justified at all, irrespective of particular empirical constraints.

6 This aspect of transcendentalism, and of internalism in particular, was elegantly ridiculed by Nietzsche in *Beyond good and evil* (I, 11): “Wie sind synthetische Urteile a priori möglich?
fragte sich Kant,— und was antwortete er eigentlich? Vermöge eines Vermögens: leider aber
nicht mit drei Worten...” (Kant put the question, How are synthetic judgments a priori possible? And what did he answer? By a capacity: but unfortunately not in three words...)

7 Other possible alternatives to internalism and externalism may be sought in the specifically social and linguistic conditions required for knowledge. In my opinion, however, these conditions are only subsidiary. I think they should be studied as structural aspects of the external world to which the individual’s cognitive structure is responsive. See also my remarks on teleology in a human setting (chapter seven).

8 The relevant passage referred to in the text is worth quoting in full, as it contains some of the same ambiguities we saw earlier in Fodor’s theory of narrow content.

“There is something strange and even absurd (etwas Befremdliches und sogar Widersinnisches) in the assertion that there should be a concept which possesses a meaning and yet is not capable of any explanation. But the categories have this peculiar feature, that only in virtue of the general condition of sensibility can they possess a determinate meaning and relation to any object. Now when this condition has been omitted from the pure category, it can contain nothing but the logical function for bringing the manifold under a concept. By means of this function or form of the concept, thus taken by itself, we cannot in any way know and distinguish what object comes under it, since we have abstracted from the sensible condition through which alone objects can come under it. Consequently, the categories require, in addition to the pure concept of understanding, determinations of their application to sensibility in general (schemata). Apart from such application they are not concepts through which an object is known and distinguished from others, but only so many modes of thinking an object for possible intuitions, and of giving it meaning, under the requisite further conditions, in conformity with some function of the understanding, that is, of defining it. (…) The pure categories are nothing but representations of things in general, so far as the manifold of their intuition must be thought through one or other of these logical functions” (Kant 1781, A 244-245).

Kant’s claim that pure categories have a meaning of their own that cannot be explained is strikingly similar to Fodor’s claim that “narrow content is radically inexpressible, because it’s only content potentially; it’s what gets to be content when—and only when—it gets to be anchored” (Fodor 1987, 50). The correspondence between Fodor’s and Kant’s views of mental content goes a long way. In the following passage from the Critique, Kant anticipated both Fodor’s notion of narrow content (reine Bedeutung), and his notion of anchoring conditions, which in Kant’s hands becomes a matter of restricting or extending (amplifizieren) a concept’s scope.

“But it is also evident that although the schemata of sensibility first realise the categories, they at the same time restrict them, that is, limit them to conditions which lie outside the understanding, and are due to sensibility. The schema is, properly, only the phenomenon, or sensible concept, of an object in agreement with the category. (…) If we omit a restricting condition, we would seem to extend the scope (amplifizieren) of the concept that was previously limited. Arguing from this assumed fact, we conclude that the categories in their pure significance (reine Bedeutung), apart from all conditions of sensibility, ought to apply to things in general, as they are, and not, like the schemata, represent them only as they appear. They ought, we conclude, to possess a meaning independent of all schemata, and of much wider application. Now there certainly does
remain in the pure concepts of understanding, even after the elimination of every sensible condition, a meaning; but it is purely logical, signifying only the bare unity of the representations. The pure concepts can find no object, and so can acquire no meaning which might yield a concept of some object. (...) The categories, therefore, without schemata, are merely functions of the understanding for concepts; and represent no object. This meaning they acquire from sensibility, which realises the understanding in the very process of restricting it” (Kant 1781, A 146-147, B 185-187)

9 A dramatic example is W.V. Quine. In spite of his professed naturalism, his semantic and epistemic holism (cf. Fodor and Lepore 1992), and his criticism of reductionism (Quine 1953), he looks upon stimuli as the evidence on which our view of the world is based. As a result of this, there is a curious asymmetry between his generally naturalist outlook and his internalist epistemology. In the following passage, for example, Quine responds to a suggestion by Donald Davidson, to the effect that ‘sameness of stimulation’, for semantic and epistemic purposes, should perhaps be identified in terms of distal factors.

“But I remain unswerved in locating stimulation at the neural input, for my interest is epistemological, however naturalized. I am interested in the flow of evidence from the triggering of the senses to the pronouncements of science. My naturalism does allow me free reference to nerve endings, rabbits, and other physical objects, but my epistemology permits the subject no such starting point. His reification of rabbits and the like is for me part of the plot, not to be passed over as part of the setting” (Quine 1990, 41-42; italics in original).

Quine the naturalist may refer to rabbits, but Quine the Cartesian remains barred behind nerves.
Bibliography


Bibliography


Chisholm, R., 1957, Perceiving: A philosophical study. Ithaca, Cornell UP.


—, 1992a, A deeper unity. Some Feyerabendian themes in neurocomputational form. In:
— and Patricia S. Churchland, 1983, Stalking the wild epistemic engine. Nous 17, pp. 5-18.
Comte, A., 1830, Cours de philosophie positive. Six volumes, Paris, 1830-1842.
—, 1991b, Mother Nature versus the walking encyclopedia. In: Ramsey, Stich, and
—, 1991, Why Fodor can’t have it both ways. In: Loewer and Rey 1991a, pp. 95-118.
—, 1985a, Fodor’s guide to mental representation. Or: the intelligent Auntie’s vademecum.
Mind 94, pp. 76-100. Reprinted in Fodor 1990a, pp. 3-30.
—, 1985b, Narrow content and meaning holism. Unpublished manuscript, MIT.
—, 1986, Why paramecia don’t have mental representations. Midwest Studies in Philosophy 10, pp. 3-23.
Goschke, Th., and D. Koppelberg, 1990, Connectionism and the semantic content of internal representation. Revue Internationale de Philosophie 172, pp. 87–103.
—, and T. Horgan, 1988, How to be realistic about folk psychology. Philosophical Psychology 1, pp. 69-81.
Hanson, N.R., 1958, Patterns of discovery. Cambridge, Cambridge UP.
Kant, I., 1781 (A), 1787 (B), Critique of pure reason. (Translated by Norman Kemp Smith, 1929, 1933, 1986. London, Macmillan.)
—, 1993, Supervenience and mind. Selected philosophical papers. Cambridge, Cambridge UP.
Mandelbaum, M., 1964, Locke’s realism. In: *Philosophy, science, and sense perception. Historical*
and critical studies. Baltimore, Johns Hopkins Press.
—, 1987, The not so happy story of the marriage of linguistics and psychology, or, Why linguistics has discouraged psychology’s recent advances. Synthese 72, pp. 341-353.
—, 1993, White Queen psychology and other essays for Alice. Cambridge, MA, MIT Press.
Moore, G.E. 1922, Philosophical studies.
Nagel, T., 1974, What it is like to be a bat. Philosophical Review 83, pp. 435-450.
Bibliography

Hillsdale, NJ, Lawrence Erlbaum.
Richardson, R., 1979, Functionalism and reductionism. Philosophy of Science 46, pp. 533-558.
Robinson, W.S., 1985, Toward eliminating Churchland’s eliminationism. Philosophical Topics 13, pp. 61-68.
— 1979, Philosophy and the mirror of nature. Princeton, Princeton UP.


300  Bibliography


aboutness 58, 164, see also: intentionality
Abrahamsen, Adele 88, 95, 97, 100, 121, 275
absorption spectra, retinal — 137–138, 279
abstract phrase structure grammar 16
access to reality 271
Ackerman, Diane 122, 278
activation space 159
actualist theories of representation 210–211, 282
Adams, R.M. 245
affections, Kant on — 247–249
affordances 262–263
AI, artificial intelligence, see also: GOFAI
Heideggerian — 255
strong — 26, 163
weak — 26
ampliative inference 106
Anderson
J.A. 275
John R. 38
annealing 89
anomalous
monism, see: token identity
trichromacy 139
antinomy of pure reason 28
Apollinaire, Guillaume 18
appereception, transcendental unity of the — 258
Aquinas, Thomas 279
arc-pair grammar 16
Aristotelian
philosophy 9, 156, 239–240, 249–252, 263
psychology and semantics 239–242
Aristotle 40, 238–242, 246, 250, 256–257, 272, 279, 285
artificial intelligence, see: AI
aspectual inspecificity 232
association
instant — 88
linear and nonlinear — 86–88, 266
pattern — 85
asymmetrical dependency 220–222, 235, 284
autoassociative memory, see: content-addressable memory
autoconnectivity 82
autonomous
behavioral descriptions 172, 177–183
content 158
psychology 168, 281
autopoeisis 93
axons 81
backpropagation 97
Baker, Lynn Rudder 279, 280, 283
balanced account of content 9–18, 261–263, 264–268, see also: dual aspect theory
Ballard, D.H. 103
Bardot, Brigitte 176
Barlow, Horace B. 70, 183
Baudelaire, Charles 176
Bechtel, William 88, 95, 97, 100, 121, 273, 275
behavior, discriminatory — 146–149
behavioral
descriptions, autonomous — 172, 177–183
dispositions 8, 18–19, 147–149, 230
language as a — disposition 11
behaviorism 12, 18–21, 272
logical — 18
methodological — 18
Bergson, Henri 57
Berkeley, George 122, 245–246
biological naturalism 29
biology, evolutionary — 231
biosemantics 12
Bliss, Timothy 276
Block, Ned 20, 26–27, 158, 163, 190, 197–199, 277, 279, 281–282, 285
Bloom, F.E. 71
Bloomfield, Leonard 18
Blumenbach, J.F. 57
Boden, Margaret 227, 275, 279
Boethius 241, 285
Bogdan, Radu 273, 279
Boltzmann, Ludwig 89–90
Boltzmann machines 89–90
bottom-up approach 12, 46, 62–70, 73, 95, 111, 150, 267
broad content 203–237
Broca, Paul 156
Bruce, V. 68, 72
Burge, Tyler 165, 168–169, 275, 279, 281–282
Butler, Charles 82
Bynum, W.F. 50, 57
calibrational content 158, 225, 281
Campbell
Donald T. 231
K. 279
Carlson, Greg 37
Carnap, Rudolf 171
Carr, Brian 39
Cartesian
metaphysics 158, 186, 189, 194–195, 236, 242–245
subject 11, 189–190, 194–195, 207
Theatre, Myth of the — 201, 243, 274, 280
categorial analysis, metaphysics as — 39–41
categories of the understanding 247–248, 258–260, 286
Caudill, Maureen 82
causal
internalism and — explanation 175–177
theories of representation 13, 208–237
center of narrative gravity 8
central
processes 35–37, 44, 282
state materialism, see: type identity theory
Changeux, Jean-Pierre 231, 276
Chapman, D. 255
Chinese Room 12, 28, 200
Chisholm, Roderick 190, 282
Chomsky, Noam 16, 19, 34
chromatic qualia 137–140
Churchland
Clark
Andrew 10, 27, 30, 79, 101, 121, 192, 272, 275–276
Austen 278
H.H. 17
Clinton
Bill 61
Mrs. 142–143
cluster analysis 101
cognition
disembodied — 30–31
explananda of — 53–57, 61, 67, 68, 72, 74, 80, 111, 162, 164, 166, 182, 197
cognitive
ethology 231
opacity 97, 110
revolution 79
science, conceptual universe for — 41–42
utility 228–237
Cohen, S.M. 273
color cube 139
color-blindness 139
combinatorial syntax and semantics 114–116
competence and performance 16–17
completion
pattern — 85
vector — 84–86, 88, 99, 106
complex, cortical — and hypercomplex cells 69, 72
computation, natural (neural) — 51, 79
computational
content 158
theory, Marr on — 68, 74
computationalism 12–13, 15, 25–26, 28, 157–200, 207
Comte, Auguste 49–50
concepts 159–161, see also: content, representation
conceptual
exploitation 123–125, 130, 145–151
interaction 12, 54–56, 64, 74, 81, 95, 112, 267
redeployment 108–109
role semantics 172, 190–192
schemes 40–41, 249
universe for cognitive science 41–42
universe for connectionist theory 92–93
Connection Machine 79, 280
connection weights, see: connectivity matrix
collectionism 11–12, 32–33, 45, 51, 56, 72, 74, 78, 79–121, 183, 267, 274
and mental content 113–121
eliminative — 33, see also: eliminativism
implementational — 33, 118–119
connectionist
models of stereoptics 72
theory, conceptual universe for — 92–93
constituent structure 114–116
constraint satisfaction 91
content 157–200, see also: representation
balanced account of — 9–18, 261–263, 264–268, see also: dual aspect theory
broad — 203–237
calibrational — 158, 225, 281
cognitive — as intrinsic 171–200
computational — 158
connectionism and mental — 113–121
correlational — 158
distributed — 93, 99–102, 158, 275
dualism of scheme and — 99
ecological — 158
formal — 158
functional — 158
hybrid explanations of mental — 254, 260–261
local — 158
opaque — 157
qualitative — of sensations 122–156
remotely controlled — 236–237, 268
strict — 158
superstrong — 11, 129–130, 133–134, 140, 142, 149–150, 157, 197, 201, 267
superweak — 11, 130, 140, 197, 267
symbolic — 158
syntactic — 158, 161
taxonomy of — 157–200
teleological — 158
translational — 158, 281
transparent — 157, 281
wide — 157, 167
content-addressable memory 88, 103
Copernican revolution 246
Copernicus 246
Corbey, Raymond 127, 152
corpuscularism 64, 242–245
correlational content 158
counterfactual theories of representation 212–213, 282
covariational theories of representation 212–213, 283
Cram, H.-R. 283
Crude Causal Theory 212–213
Cussins, Adrian 275
Dalton 65
Darth Vader 58, 77, 190–191
Darwin, Charles 234
Davidson, Donald 21, 281–282, 287
deduction
empirical — 285
metaphysical — 285
transcendental — 13, 238, 252–254, 285
degradation, graceful — 103
deictic representation 255
Dell, Gary S. 37
delta-rule 97
demonology 24
dependency, asymmetrical — 220–222, 235, 284
descriptive metaphysics  272

design stance  27
Devitt, Michael  161, 199, 279
dichromacy  139
Ding an sich  248, 258
direct
  acquaintance  205–206
  perception  153, 263–264
  pickup of invariant properties  261–263
discriminatory behavior  146–149
disembodied, mind as a — entity  30–31, 186
disjunction problem  214–216, 283, see also: misrepresentation
dispositions, behavioral —  8, 18–19, 147–149, 230
distal inspecificity  232
distributed representation  93, 99–102, 158, 275
dot product, see: inner product
double aspect theory  77
Double, R.  273
Dretske, Fred I.  123, 192, 212, 216–219, 222, 233, 279, 281–284
Dreyfus, Hubert L.  85, 255, 276
Driesch, Hans  57
dual aspect theory of folk psychology  12, 46, 75, 77, 81, 156, 203–204
dualism  20, 40, 236
  of scheme and content  249
dynamics, network —  89, 267

dynamical
  content  158
  psychology  231, 238
  realism  255
Edelman, G.M.  231, 276
Einstein, Albert  27
eliminative materialism, see: eliminativism
eliminativism  10, 12, 31–34, 41–42, 45–78, 93, 118–119, 122–124, 129, 133, 137, 141, 155, 162, 166, 179, 181–184, 267, 272–273
Eliot, Lewis  274
elliptic nature of folk psychology  53
embodiment, conditions of —  30–31, 186
emergent properties  117–118
empathy, intersubjective —  136–137, 150
empirical
  deduction  285
knowledge, unmoved movers of —  7
realism  126–127, 132
empiricism  7, 9, 94–98, 101, 126–127, 129, 132, 151, 277
empiricist nets  110, see also: tabula rasa
Enc, Berent  65
encapsulation, informational —  34–36, 98, 279
epiphenomenalism  133, 137
epistemology, neural —  12, 33, 79–121, 267–269
epistemological hylomorphism  249–252
epistemology  9, 172, 207, 239, 247, 271
  evolutionary —  231
  internalism and —  193–195
  metaphysics and —  9, 239–258
  naturalized —  256, 267, 271
equipotentiality, see: Mass Action
essentialism  176
ethology, cognitive —  231
evidential interaction, see: conceptual interaction
  evolutionary
    biology  231
    epistemology  231
  experiential realism  254
  explananda of cognition  53–57, 61, 67–68, 72, 74, 80, 111, 162, 164, 166, 182, 197
  exploitation, conceptual —  123–125, 130, 145–151
  exterospection  143–144, 154–156

faculty
  of sensibility  248
  of understanding  248
  psychology  34

feature detection model  70–71, 82–85, 111
feedforward networks  81–83
Feldman, J.A.  103
Feyerabend, Paul K.  45, 47, 108, 150, 279
finite state machine  19, 264
Fischler, O.  16, 138
First Law, Fodor’s — of the Nonexistence of Cognitive Science  37, 44
first philosophy  9, 239, 247, 271
Fischler, M.A.  16, 138
Flores, Fernando  255
Index


Fodor-Stich paradox  197–200

folk biology  55, 59

descriptive vocabulary of —  46, 66
dual aspect theory of —  12, 46, 75, 77, 81, 156, 203–204
elliptic nature of —  53
fallibility of —  46

formal
content  158
properties  162–164

formality condition  171, 186–190, 281
foundationalism  10, 56
frame problem  36
Freud, Sigmund  191
Frisby, John P.  68, 281

function, natural —  222, 224–237

functional
content  158
misrepresentation  228–229

narrow —  185, 281
wide —  168, 185–186, 281

Galileo  27
Gall, Franz Joseph  34
ganglion cells, retinal —  68–71
Gardner, Howard  15, 20, 79, 98
Garfield, Jay  37, 169, 181, 193, 273, 275, 279, 282
Garon, J.  79, 181
generative semantics  16
Gibson, J.J.  154, 194, 202, 205, 231, 238, 255, 261–264, 282
given  7, 10–11, 40, 129, 134, 151, 156, 196–197, 201, 206, 236, 259, 265, see also: intrinsic content
Myth of the —  7–10, 13, 236, 268
God’s eye view  256–260

GOFAI  80, 164, see also: AI
Goldman, Alvin I.  273, 279
Goodman, Nelson  234, 235
Gopnik, I. and M.  37, 275
Goschke, Thomas  275
Gould, Stephen Jay  104
government and binding theory  16
graceful degradation  103
grades of theoreticity  273
Graham, G.  273

grammar  15–19, 33
abstract phase structure —  16
arc-pair —  16
generative semantics and —  16
government and binding theory and —  16
realistic —  16
standard theory and —  16
syntactic structures and —  16

Green, P.  68, 72
Greenwood, J.D.  273
Gregory, Richard  15, 36, 139, 151, 278
Gulliver  14–15, 34
Gunderson, Keith  170

Hamlyn, D.W.  39, 240
Hanson, N.R.  108, 151
Hare, R.M.  281
Harnad, Stevan  140
Harrison, Harry  276
Hartline, P.H.  131, 226
Hatfield, Gary  275
Haugeland, John  26, 58, 80, 158, 163–164, 173
Haviland, S.E.  17
Hawkins, R.D.  276
Heidegger, Martin  255
Heideggerian AI  255
Heil, John  168, 173, 279
hidden nodes  81, 87
Hillis, W. Daniel  79, 280
Hinton diagrams  101–102, 110, 117
Hinton, Geoffrey E.  88, 90, 97, 100, 102, 121, 275–276
Hofstadter, Douglas R.  27, 120, 275, 277
homunculi  22–25, 187
homunculus fallacy  23
Hopfield, John  89, 91, 275
Hopfield net  89–90
Index

Horgan, Terence  273
horizontal faculty psychology  34
Hubel, David  68–71
Hume, David  245–247
Husserl, Edmund  281
Huygens, Christian  108, 109
hybrid explanations of mental content  254, 260–261
hylomorphism  239–242, 254
epistemological —  249–252
metaphysical —  249–250
hypercortex, cortical complex and — cells  69, 72

idealism
refutation of —  155, 279
transcendental —  257–260

identity theory, mind-brain — , see also: physicalism, reductionism
token —  21, 173–175
type —  18–21, 41, 43, 173
illusions, optical —  35–36
implementational connectionism  33, 118–119
incommensurability  51, 57, 74, 78, 80, 108
indicator theories of representation  212–213
individualism  165, 168
methodological —  175–177
induction, new riddle of —  234
ineffability  122, 136–137, 150
informational encapsulation  34–36, 98, 279
innate ideas  243, 283, see also: nativism
inner product  83–84
inspecificity
aspectsual —  232
distal —  232
linear —  232–233
representational —  173, 231–236
instant association  88
intentional stance  27, 59
intentionality  12, 15, 28–31, 43, 50–52, 57–61, 74, 167, 236, see also: aboutness
interaction
with the environment  203–237, 223–237, 254
conceptual —  12, 54–56, 64, 74, 81, 95, 112, 267
interactionist programming  255
internal
realism  168, 238, 255–260
representation  11, 60, 78
subject  235–236
internalism  13, 80, 157–200, 202, 223, 238–269, 281, 285–286
and autonomous behavioral descriptions  177–183
and causal explanation  175–177
and computationalism  186–190
and conceptual role semantics  190–192
and epistemology  193–195
and externalism  167–171, 251–254
and functionalism  184–186
and neurophysiology  183–184
and scientific methodology  192–193
and supervenience  173–175
intersubjective empathy  136–137
see also: given
introspection  143–144, 153–156
intuitions
about the mental  27, see also: surrealism
do sensitivity  247–248
invariant properties  261–263
inversion of qualia  136–137, 150
isotropic and Quinean nature of central processes  37, 282
Jackson, Frank  148, 273, 277, 279
Jaynes, Julian  201
Jerison, H.J. and I.  231, 276
Johnson, Mark  198, 255
Johnson-Laird, Philip N.  16, 97
Jordan, M.I.  82
judgmental plasticity, see: weak plasticity
Kandel, E.R.  276
Kanizsa, Gaetano  36
Katz, Jerrold  17
Kim, Jaegwon  173–174, 281
Kinsbourne, Marcel  280
Kitcher, Patricia  168–169, 185, 188, 193, 273, 275, 279, 281
Klahr, D.  275
knowledge
as an empirical phenomenon  7–13
as radically linguistic  7
as real  7–13, 259
balanced account of —  9–18, 261–263, 264–268, see also: dual aspect theory
social constraints on —  11, 230, 286
unreality of —  7–13
Koenderink, J.J.  205, 263
Koppelberg, Dirk  275
Kosslyn, Stephen  274
Kretzmann, Norman  272
Kripke, Saul A.  19, 164, 210–211, 277, 282
Kuffler, Stephen W.  68, 71
Kuhn, Thomas S.  79–80
labeling fallacy  119–121, 140–142
Lakatos, Imre  41, 46, 50, 53, 92, 273
Lakoff, George  31, 38, 164, 176, 198, 254–255, 281, 283
Land, Edwin  139, 278
language
and cognition  230
as a behavioral disposition  11
as an abstract Platonic object  17
of thought  60, 115–116
organ  34, 38
processing  22, 24, 32
Lashley, Karl  19, 20, 98
lateral geniculate body  71, 73
Laudan, Larry  46
Lavoisier, A.-L.  65
learning  96, 276
Lehky, S.R.  72, 74
Leibniz, G.W.  190, 200, 244
Lepore, Ernest  129, 137, 140–141, 191, 279, 282, 287
Lettvin, J.Y.  232
Levelt, Wim  22
Lewis, David  282
Lewontin, R.  104
Lincoln, Abraham  142–143
Lindsay, Peter  23–24
linear
and nonlinear association  86–88
inspecificity  232–233
linguistic
catastrophe  60
knowledge as radically —  7–13
plasticity, see: weak plasticity
theory  15–17
turn  8–9, 249
Lloyd, Dan  102, 192, 233, 276, 283
local content  158
localist models  99, 101
Locke, John  95, 129, 132, 223, 244–245, 250, 277–278
Loewer, Barry  157, 199, 204, 209–210, 212, 275, 279–281, 283
logical behaviorism  18
Luria, Aleksandr R.  35, 274, 276
Lycan, William G.  18, 62
Lyons, J.  275
W.  279
Lømo, Terje  276
Madell, G.  273
Malebranche, Nicholas  244
Mandelbaum, Maurice  278
manifest image  242
mapping, pattern —  85–86, 266
Margolis, H.  85
Marr, David  16, 34, 68, 70, 75, 95, 169, 275
Marslen-Wilson, William  38
master cards  159–165, 171
materialism  272
central state — , see: type identity theory
eliminative — , see: eliminativism
reductive — , see: type identity theory
matrix
connectivity —  81, 85–86, 99, 101, 106
multiplication  85–88
McCauley, R.N.  16, 38, 273
McClelland, Jay L.  32, 51, 79, 90–91, 94–95, 97, 100, 106, 275
McGinn, Colin  168, 279, 283
Medin, D.L.  275, 281
Meijering, Theo  37, 239, 242, 279
Mejsing, Monica  78
mental content, see: content, representation
metaphysical
hylomorphism  249–250
deduction  285
dualism  236
realism  164, 168, 242, 255–260
metaphysics  9, 12, 39, 41, 44, 239–258, 271
aim of —  41
and epistemology 9, 239–258
as categorial analysis 39–43
descriptive — 272
naturalized — 39–43
revisionary — 272
methodical doubt 243
methodological
behaviorism 18
individualism 175–177
solipsism 168, 171–173, 281, see also: internalism
methodology
bottom-up vs. top-down — 63–73
internalism and scientific — 192–193
microfeature analysis 99–102, 111–112, 117, 275
microfunctionalism 30, 38, 43
micromodularity 38, 101
Millikan, Ruth Garret 12, 184, 213, 229, 235, 254, 279, 282, 284
mind-brain identity, see: identity theory
Minsky, Marvin 227, 276
misrepresentation 173, 214–222, 282–283
Dretske on — 216–219
Fodor on — 219–222
functional — 228–229
modularity 12, 34–36, 38, 41–44, 93, 152–153, 275, 279
and functional holism 98–99
in natural language processing 37
monism, anomalous —, see: token identity theory
Montague, R. 16
Moore, G.E. 281
Morris, W.E. 282, 283
Mountcastle, V.B. 276
Müller-Lyer, F.C. 35, 36
multiple realization 43, 65, 174, 185–186
multiplication, matrix — 85–88
Murre, Jaap 98
Myth
of the Cartesian Theatre 201, 243, 274
of the Given, see: given
Nagel, Thomas 277
narrative gravity, center of — 8
narrow
content 157, 167, 197–200, 203–204, 222, 260–261, 281, 286
functionalism 185, 281
nativism 94–98, 101, see also: innate ideas
natural
computation 51, 79–121
epipistemology 33, 79–121
function 222, 224–237
kinds 164, 176, 210
teleology 224–238, 254, 268
naturalism 7–13, 17, 39–40, 168, 271, 287
biological — 29
naturalistic
epipistemology 256, 267, 271
fallacy 7–13, 129, 130, 157, 196–197, 236, 268
metaphysics 39, 41, 43
psychology 192
Necker cube 90, 91–95, 100, 106
Neisser, Ulric 194, 231, 282
network
dynamics 89, 267
semantics 132
theory of meaning 126–127
topologies 82
neural
computation 10, 79
epipistemology 12, 79, 267–269
neurophysiology, see also: eliminativism
bottom-up approach in — 63–73
internalism and — 183–184
New Look theories of perception 150, 153
new riddle of induction 234
Newell, Allen 158, 163
Newman, E.A. 131, 226
Nietzsche, Friedrich 285
non-individualism 168
non-linear transfer functions 87, 266
nonlinear association 87, 266
nonrelational taxonomy of properties 175–177
Norman, D. 23, 24, 80
noumenal world, see: Ding an sich
observation 249
as theoryladen 150
sentences 146–149, 274, 277
observationality
of folk psychology 54, 56, 61, 74, 182
degree of — 53, 273–274
one hundred step rule 103
ontological hylomorphism 249–252
opacity
cognitive and semantic — 97, 110
referential — 190, 282
opaque
cognitive and semantic — 157, 281
taxonomy 281
optical illusions 35–36
ordo
cognoscendi 254
essendi 254
Ovid 271
Owens, J. 279
Pandemonium 275
Papineau, David 173, 213, 229, 279, 282, 284
parallel distributed processing 32, 51, 79, see also: connectionism
partitioning of state space 107–110, 117, 146–150, 152–153
Pascal, Blaise 67, 274
past tense formation of English verbs 94–98
pattern mapping 85–86, 266
Peacocke, Christopher 205
Peirce, Charles S. 67
percept 159–161
Perceptron 275
perceptual grid 147–149
performance, competence and — 16–17
Peterson, Roland 275
Pettit, Philip 273
phenomenal world 248
Philipse, Herman 127–133, 142–144, 154
philosophical surrealism 27, 277–278
phenomenology 34
physical stance 27
physicalism 13, 20, 173, 175, 207, 277, 279, see also: identity theory
pineal gland 243
Pinker, Steven 33, 97, 275
plasticity
and learning 93–98, 103
of mind 19, 93, 98, 105, 110, 125–129, 229, 268, 275, 276
of perception 125–129
judgmental —, see: weak plasticity
linguistic —, see: weak plasticity
sensational —, see: strong plasticity
strong — 128, 130, 142–150, 153–156
weak — 128, 131–132, 134
Plato 17, 156
Poggio, Tomaso 169
Pols, E. 8, 278
Ponzo 36
positivism 49
possible worlds 164, 210–211, 227
Possin, Kevin 172–173, 279, 282–283
Poulenc, Francis 18, 19
Präformationssystem 244, 258
Preston, J.M. 273
primary
qualities 244–245
visual cortex 69
Prince, A. 33, 97, 275
Principle of Mass Action 19–20, 98
projective fields 73
propositional attitudes 47, 50, 58, 60–61, 68, 74–78, 155–156, 164, 204, 281
dual aspect theory of the — 12, 46, 75, 77, 81, 156, 203–204
elliptic nature of the — 53
prototype vector 88, 107–110, 159
psycholinguistics 21–22, 38
psychology
ecological — 231, 238
faculty — 34
psychosemantics 12
punch card 158–161, 187–190
pure manifold of affections 249–250
Purkinje cell 105
Purves, D. 276
Pylyshyn, Zenon W. 9, 26, 30, 34, 41–42, 60, 78, 80, 114, 117–121, 167, 205, 263–264, 275, 277, 280
qua-problem 211, 232
qualia 13, 122–156, 277, 278, 279
and epistemic content 122–156
chromatic — 137–140
inversion of — 136–137, 150
strong — 130, 144–149, 146, 149–150
superstrong — 133–134, 142, 149–150
superweak — 130, 133–134
weak — 130–131, 133–136, 149–150
qualitative content of sensations 122–156
Index

Quine 8, 21, 37, 47, 53, 123, 126, 271, 273–274, 277, 282, 287
Quinean and isotropic character of central processes 37, 282

Ramsey, W. 79, 181, 273
ready-made
  categorial structure 40
  world 242

realism
  ecological — 255
  empirical — 126–127, 132
  experiential — 254
  internal — 168, 238, 255–260
  metaphysical — 164, 168, 242, 255–260
  scientific — 126–127
realistic grammar 16
receptive field 68–73
recognition, pattern — 85
recurrent net 82
redemption, conceptual — 108–109
reductionism 10, 12, 19, 41, 54–56, 124–125, 129, 141, 173, 259, 271, 278, 287, see also:
  type identity theory
reductive materialism, see: type identity theory
referential opacity 190, 282
refutation of idealism 155, 279
relational
  nature of cognitive states 202–204, 207
  properties 175
relationism 13, 238, 253–254
relaxation net 89, 91–93
remotely controlled content 236–237, 268
replacement argument 178, 182, 282
representation, see also: content
  actualist theories of — 210–211, 282
  causal theories of — 13, 208–237
  counterfactual theories of — 212–213, 282
  covariational theories of — 212–213, 283
  Crude Causal Theory of — 212–213
decitic — 255
distributed — 93, 99–102, 158, 275
indicator theories of — 212–213
representational specificity 173, 231–236
representationalism 11
res
  cogitans 236, 243
  extensa 236, 243
Rescher, Nicholas 48
retinal
  absorption spectra 137–138, 279
  ganglion cells 68–71
  retinex theory 139, 279
revisionary metaphysics 272
revisionism 33, 118–119
revolution
  cognitive — 79
  Copernican — 246
  scientific — 79
Rey, Georges 157, 199, 204, 209–210, 212, 275, 279–281, 283
Richards, W. 51, 79, 85, 140
Richardson, R. 65
rigid designators 164, 210–211
Robinson
  H. 277
  W.S. 273
robust tokens 215–216, 283
robustness 103
Rock, I. 151
Rorty, Richard 8–9, 11, 18, 45, 151–152, 193, 196–197, 243, 257
Rosch, Eleanor 281
Rosenberg, C. 101
Rosenfeld, E. 275
Ross, D. 37, 279
Roth, Ilona 68, 281
rules and representations 116–118
Rumelhart, David 32, 51, 79, 90–91, 94–95, 97, 100, 106, 121, 275
Russell, Bertrand 47, 205
Ryle, Gilbert 18, 201, 243, 274

scalar product, see: inner product
scheme
  conceptual — 40–41, 249
  dualism of — and content 249
scholasticism 158
Schwartz, Robert 246, 264
scientific
  image 242
  internalism and — methodology 192–193
realism 126–127
revolution 79
Searle, John R. 12, 18, 26–31, 43, 163, 200, 259, 268, 277, 282
secondary qualities 244–245
Seidenberg 37
Sellars, Wilfrid 7–13, 47–48, 129–130, 242, 259, 268, 278
semantic opacity 97, 110
transparency 101
semantics conceptual role — 172, 190–192
generative — 16
sensational plasticity, see: strong plasticity
sensations 122–156
sensibility, faculty of — 248
sensory impressions 8
sensus communis 241
sententialism 60, 74–75, 274
serendipity 108
Shakespeare, William 104
Shannon, Claude 216, 283
Sharpe, R. 273
Simon, Herbert A. 158, 163
simple, cortical — cell 69–72
simplicity 108
skepticism 48
Skinner, B. Frederic 18–19
Skywalker, Luke 58, 77, 190–191
Sleutels, Jan 127, 152
Smith Adam 260
E.E. 275, 281
Smolensky, Paul 51, 79, 100, 120–121, 167, 275–277
Snow, C.P. 274
social constraints on knowledge 11, 230, 286
society of mind 227, 276
soft constraints 103
solipsism 165
methodological — 168, 171–173, 281, see also: internalism
special sciences 21, 176
species chauvinism 20
specificity, representational — 173, 231–236, see also: inspecificity
Spinoza, Baruch 244
Spruit, Leen 239
Stalnaker, Robert C. 212, 283
Stampe, Dennis W. 212, 279, 282
standard theory 16
state space, partitioning of — 107–110, 117, 146–150, 152–153
Sterelny, Kim 169, 181, 209, 211–212, 219, 228, 234, 273, 279, 281, 283
stereoptics, connectionist models of — 72
stereotypes 272
Stillings, Neil 168–169, 275, 281, 283
Strawson, Peter 78, 272
strict content 158
strong AI 26, 163
plasticity 128, 130, 142–150, 153–156
qualia 130, 144–149, 146, 149–150
subcomputational properties 161–164
subject Cartesian — 11, 189–190, 194–195, 207
internal — 235–236
subjectivity 136–137, 150, 201
Cartesian metaphysics of — 158, 194–195
ineffable — of qualia 122
subsymbolism 11, 277
supermatrices 105
superstrong content 11, 129–130, 133–134, 140, 142, 149–150, 157, 197, 201, 267
qualia 133–134, 142, 149–150
supervectors 105
supervenience 163, 172–175, 206, 281
superweak content 11, 130, 140, 197, 267
qualia 130, 133–134
surrealism, philosophical — 27, 277–278
Swift, Jonathan 14, 18
symbolic content 158
symbolism 11
synapses 81
syntactic content 158, 161
engine 115, 163
properties 162–164, 187–190
structures 16
theory of mind 163
syntax, combinatorial — and semantics 114–116
synthetic judgments a priori 286
Szentágothai, J. 276

*tabula rasa* 94, 275
Tanenhaus, Michael K. 37
taxonomy
  of content 157–200
  of sensations 124–125
  opaque — 281
Taylor, K.A. 228, 280, 282–283
teleological
  content 158
teleology, natural — 224–238, 254, 268
theoreticity, grades of — 273, see also: observationality
theory
  and observation 150–153
  of content 157–200
theory-ladenness of observation 108, 126–127, 150, 153, 279
thermodynamic systems 89–90
third-person explanations 123
Thomas Aquinas 279
thought experiments, see: surrealism
three stage doctrine 49
token-physicalism, see: token identity theory
tokens
  robust — 215–216, 283
  wild — 215–216, 283
top-down approach 12, 6–64, 67, 74, 95, 111, 150, 186, 267, 274
topologies, network — 82
Touretzky, D.S. 121
transcendental
  arguments 28, 195, 227, 248
  deduction 13, 238, 252–254, 285
  features of knowledge 28, 40, 285
  idealism 257–260
  philosophy 231, 237, 245–249, 247, 251–252
  unity of the apperception 258
transducers 136
transfer function 86–88

non-linear — 87, 266
transformation, pattern — 85
translational content 158, 281
transparency, semantic — 101
transparent content 157, 281
Turing machine 92, 158
Turing, Alan M. 118
Twin-Earth 164–167, 174, 178, 210, 282
Tyler, Lorraine 38
type-physicalism, see: type identity theory
über-sätzenal attitudes 155, 156
Ullman, Shimon 263
understanding, faculty of — 248
unitary view of cognitive architecture 34, 38
unmoved movers of empirical knowledge 7
unreality of knowledge 7
utility, cognitive — structure 223–237, 253–254, 284

Vader, Darth 58, 77, 190–191
Van Gelder, Tim 98, 276
Van Helmont, J.B. 57
vector
  coding 104–110, 111, 137, 276
  completion 88, 106
  dynamics 267
  inner product of —s 82–85
  transformation 85–86, 104–110
vertical faculty psychology 34
*virtus dormitiva* 263
visual system, mammalian — 68–71
vitalism 31, 51, 57, 59
Von Neumann
  bottleneck 280
  machine 79, 92
Von Neumann, John 280

wager, Pascal’s argument of the — 67
Watson, John B. 18
Way of Ideas 205–206
weak
  AI 26
  plasticity 128, 131–132, 134
  qualia 130–131, 133–136, 149–150
Weaver, Warren 216, 283
<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight matrix, see:</td>
<td></td>
<td>wild tokens</td>
<td>215–216, 283</td>
</tr>
<tr>
<td>connectivity matrix</td>
<td></td>
<td>Wilkes, Kathleen</td>
<td>278</td>
</tr>
<tr>
<td>Wernicke, Carl</td>
<td>156</td>
<td>Williams, M.</td>
<td>283</td>
</tr>
<tr>
<td>Wickelfeature</td>
<td>100</td>
<td>Winograd, Terry</td>
<td>255</td>
</tr>
<tr>
<td>Wickelgren, W.A.</td>
<td>100</td>
<td>Wittgenstein, Ludwig</td>
<td>8</td>
</tr>
<tr>
<td>Wickelphone</td>
<td>100</td>
<td>Woodfield, Andrew</td>
<td>280</td>
</tr>
<tr>
<td>wide</td>
<td></td>
<td>Woodward, J.</td>
<td>273</td>
</tr>
<tr>
<td>content</td>
<td>157, 167</td>
<td>Yolton, J.W.</td>
<td>278</td>
</tr>
<tr>
<td>functionalism</td>
<td>168, 185–186, 281</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiesel, Torsten Nils</td>
<td>68–71</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Samenvatting

De werkelijkheid van kennis

De filosofische onwerkelijkheid van kennis

Filosofen beweren dat onze kennis van de werkelijkheid niet werkelijk is. Kennis gaat immers logisch vooraf aan werkelijkheid, en kan er dus als zodanig geen deel van uitmaken. Het is dan ook niet mogelijk om kennis te bestuderen vanuit een zogeheten ‘naturalistisch’ gezichtspunt, als deel van de wetenschappelijk beschrijfbare natuur. In deze studie wordt het tegengestelde standpunt verdedigd, namelijk dat kennis wel werkelijk is en dat zij ook als zodanig bestudeerd dient te worden. Hetzelfde geldt voor de subjecten of dragers van kennis, in het bijzonder voor menselijke wezens. Filosofen plaatsen het subject vaak buiten de gekende werkelijkheid, als een soort van bovennatuurlijke buitenstaander. Naar mijn mening is ook het subject deel van de natuur: kennis en drager zijn even werkelijk. Aangezien bij de mens, het standaardgeval van een kennend wezen, het zenuwstelsel fungeert als het organ van kennis, dient de wetenschap van werkelijke kennis een wetenschap van het brein te zijn.

Op zich lijkt de onwerkelijkheid van kennis een typisch filosofisch en weinig praktisch probleem. Toch speelt het ook in de wetenschap een rol van betekenis. Heden ten dage richten diverse wetenschappen zich uitdrukkelijk op de studie van cognitieve verschijnselen. Als verzamela naam voor deze cluster van disciplines, waartoe onder meer de psychologie en de neurofysiologie behoren, wordt de term ‘cognitiewetenschap’ gebruikt. Ofschoon de cognitiewetenschap kennis beslist wil beschouwen als een empirisch fenomeen, blijkt zij sterk te worden beïnvloed door de filosofische onwerkelijkheid van kennis. Deze invloed is merkbaar op tal van punten, waaronder de afbakening van het onderzoeksdomein, de keuze van een verklarend begrippenstelsel en de keuze van een geschikte methodologie. De deelnemende disciplines ondergaan de invloed elk op hun eigen manier. Cognitieve psychologen stellen doorgaans dat niet het feitelijke brein bestudeerd dient te worden, maar de software van het brein: een abstracte, virtuele machine, derhalve een onwerkelijkheid. Computationalisten zien kennis vooral als een kwestie van symbolen in een universele denktaal: een taal die gaat over de werkelijkheid en die zich

_Tegen gegevens_

In de hedendaagse filosofie neemt de gewraakte onwerkelijkheid van kennis steeds duidelijker gestalte aan. Een belangrijke bijdrage hiertoe werd geleverd door de zogeheten ‘linguïstische omwenteling’ in de Angelsaksische filosofie, de toenemende belangstelling voor de _taal_ waarin wij de werkelijkheid beschrijven, als onderscheiden van deze werkelijkheid zelf. Naast Wittgenstein, Goodman en Quine is vooral ook Wilfrid Sellars verantwoordelijk voor deze ontwikkeling. Als mijlpaal geldt Sellars’ kritiek op de ‘Mythe van Gegevens’ (‘Myth of the Given’). Sellars verwerpt het idee dat kennis volledig begrepen kan worden in termen van een verzameling vastliggende feiten of empirische gegevens, bijvoorbeeld zintuiglijke indrukken, mentale symbolen, gedragsdisposities, psychologische ‘software’ of hersentoestanden. Kennis bezitten is volgens Sellars niet een bepaalde empirische toestand; het is een logische en linguïstische toestand, namelijk die van opgenomen te zijn in een gemeenschap van taalgenoten die bepaalde kennisclaims aanvaarden. Kennis bezitten is deelgenoot zijn in het vertoog van een gemeenschap. Het is een filosofische mythe dat aan dit vertoog ‘gegevens’ ten grondslag zouden liggen waarin kennis wetenschappelijk verankerd kan worden. ‘Naturalisme’ in de kentheorie moet daarom worden verworpen: het vervalt onherroepelijk tot een drogredenering gebaseerd op de Mythe van Gegevens.

De gevolgen van deze omwenteling in de filosofie zijn opzienbarend. Zo wordt het subject van kennis gaandeweg getransformeerd van een empirische persoon tot een abstracte knoop in een web van vertogen. Het Ik wordt sublimieerd tot deelnemer aan de “conversatie der mensheid” of tot een “centrum van narratieve zwaartekracht”, zoals Richard Rorty en Daniel Dennett onlangs hebben betoogd. Deze verbanning van het subject uit de werkelijkheid waarover het vertoog der mensheid gaat is filosofisch gezien een groot probleem: wat is de aard van het subject, als het niet gewoon werkelijk is?

Wanneer het subject van kennis buiten de gekende werkelijkheid wordt
geplaatst, wordt ook de band tussen subject en werkelijkheid een probleem. De werkelijkheid wordt een ‘buitenwereld’, terwijl haar invloed op onze kennis ontaardt in een reeks nondescripte prikkels. Het subject creëert uit deze prikkels naar eigen inzicht een wereldbeeld. Voor de moderne, empiristisch en realistisch georiënteerde filosofie is dit verarmde contact met de werkelijkheid bepaald desastreus te noemen.

Met de verwerping van het naturalisme in de kentheorie wordt ten slotte ook kennis zelf iets onwerklijks. Kennis kan niet langer worden bestudeerd als een empirisch fenomeen. Het is een louter logisch verschijnsel, een boven-natuurlijke schakeling van vertogen als onderscheiden van de natuurlijke werkelijkheid die door deze vertogen wordt geconcipieerd. Als Sellars’ kritiek op de Mythe van Gegevens steekhoudend is, betekent dit voor de empirische cognitiewetenschap dat zij conceptueel onmogelijk is.

**Naturalisme zonder gegevens**

Sellars’ uitdaging aan het naturalisme is van fundamenteel belang. Bovengenoemde consequenties van zijn kritiek dienen echter niet minder ernstig te worden genomen. In deze studie probeer ik te komen tot een meer evenwichtige visie op kennis, waarin aan de werkelijkheid van kennis evenveel recht kan worden gedaan als aan het logische karakter ervan. Op deze manier hoop ik de weg te banen voor een empirische wetenschap van kennis die niet ten prooi valt aan Sellars’ kritiek op het naturalisme.

In hoofdzaak worden twee concurrerende onderzoeksprogramma’s in de cognitiewetenschap met elkaar vergeleken, het computationalistische programma van de cognitieve psychologie en het naturalistische programma van de cognitieve neurowetenschap. Op de keper beschouwd maken beide gebruik van de door Sellars verworpen mythe. Zij zoeken naar een klasse van empirische gegevens in termen waarvan cognitie uitputtend kan worden begrepen: computationalisten beroepen zich op mentale symbolen, naturalisten op hersentoestanden. Het voornaamste verschil tussen beide is dat symbolen vooral het logische karakter van kennis benadrukken, terwijl hersentoestanden bij uitstek werkelijk zijn. Als alternatief voor deze benaderingen argumenteer ik voor een vorm van cognitiewetenschap waarin beide aspecten tot hun recht kunnen komen. De neurocomputationele benadering die in dit kader wordt voorgesteld komt tot op zekere hoogte overeen met recente suggesties van zogenoemde ‘connectionisten’. Nieuw is echter de poging om naturalisme en computationalisme te combineren op een wijze die geen gebruik maakt van de Mythe van Gegevens.
Globaal samengevat argumenteert deze studie voor een neurale kentheorie of epistemica die geënt is op het connectionisme in de cognitiewetenschap. Na een inleidende schets van enkele hoofdproblemen uit de hedendaagse filosofie van de geest (hoofdstuk 2) wordt uitgebreid ingegaan op het eliminatief materialisme, zoals voorgesteld door de Amerikaanse filosoof Paul Churchland, en op de rol van de ‘volkspsychologie’ in de cognitiewetenschap (hoofdstuk 3). In dit verband pleit ik voor een vergaande integratie en conceptuele wisselwerking tussen psychologie en neurofysiologie.


Niet alleen qualia worden traditioneel begiftigd met intrinsieke inhoud, maar mentale toestanden in het algemeen. De semantiek van mentale toestanden komt uitgebreid aan de orde in hoofdstuk 6 en 7. Ik ga na hoe het vooroordeel dat mentale inhoud intrinsiek bepaald is verankerd ligt in traditionele vormen van fysicalisme en computationalisme, en op welke wijze het terugkeert in moderne, zogeheten ‘causale’ verklaringen van mentale representatie. In dit kader worden twee globale onderzoeksprogramma’s geïdentificeerd, ‘internalisme’ en ‘externalisme’, die de bepaling van mentale inhoud zoeken in respectievelijk het subject en zijn omgeving. In een kritiek op beide programma’s dring ik aan op een meer evenwichtige opvatting van mentale inhoud, waarin de bijdrage van zowel het subject als zijn omgeving kan worden verdisconteerd.

Hoofdstuk 8 beschouwt het probleem van mentale inhoud vanuit een breder historisch en filosofisch perspectief. Als radicaal alternatief voor het gangbare internalisme en externalisme stel ik een relationistische theorie van mentale inhoud voor. De ware kracht van het relationisme ligt in de afwijzing van elke vorm van intrinsieke inhoud. Deze heroriëntatie met betrekking tot mentale inhoud opent de weg voor een empirische cognitiewetenschap zonder intrinsieke mentale ‘gegevens’ en vrij van Sellars’ kritiek op het naturalisme.
Curriculum vitae