The Philosophy of Psychology

George Botterill

and

Peter Carruthers

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1999

Printed in the United Kingdom at the University Press, Cambridge

Typeset in Times 10/12pt [VN]

A catalogue record for this book is available from the British Library

Library of Congress cataloging in publication data

Botterill, George.
The philosophy of psychology / George Botterill, Peter Carruthers.
p. cm.
Includes bibliographical references and indexes.
ISBN 0 521 55111 0 (hardback) – ISBN 0 521 55915 4 (paperback)
1. Psychology – Philosophy. I. Carruthers, Peter, 1952– .
II. Title.
BF38.B63 1999
150′.1 – dc21 98–33301 CIP

ISBN 0 521 55111 0 hardback
ISBN 0 521 55915 4 paperback

Contents

1 Introduction: some background

Readers of this book should already have some familiarity with modern philosophy of mind, and at least a glancing acquaintance with contemporary psychology and cognitive science. (Anyone of whom this is *not* true is recommended to look at one or more of the introductions listed at the end of the chapter.) Here we shall only try to set the arguments of subsequent chapters into context by surveying – very briskly – some of the historical debates and developments which form the background to our work.

1 Developments in philosophy of mind

Philosophy of mind in the English-speaking world has been dominated by two main ambitions throughout most of the twentieth century – to avoid causal mysteries about the workings of the mind, and to meet scepticism about other minds by providing a reasonable account of what we can know, or justifiably infer, about the mental states of other people. So most work in this field has been governed by two constraints, which we will call *naturalism* and *psychological knowledge*.

According to *naturalism* human beings are complex biological organisms and as such are part of the natural order, being subject to the same laws of nature as everything else in the world. If we are going to stick to a naturalistic approach, then we cannot allow that there is anything to the mind which needs to be accounted for by invoking vital spirits, incorporeal souls, astral planes, or anything else which cannot be integrated with natural science. Amongst the thorniest questions for naturalism are whether thoughts with representational content (the so-called *intentional states* such as beliefs and desires, which have the distinctive characteristic of *being about something*), and whether experiences with phenomenal properties (which have distinctive subjective feels, and which are *like something* to undergo), are themselves suitable for integration within the corpus of scientific knowledge. We will be addressing these issues in chapters 7 and 9 respectively.

Psychological knowledge has two aspects, depending upon whether our knowledge is of other people or of ourselves. Different accounts of the mental will yield different stories about how we can have knowledge of it, or indeed whether we can have such knowledge at all. So a theory of mind ought to fit in with a reasonable view of the extent and nature of psychological knowledge. The details of the fit are a somewhat delicate matter. It must be conceded that both empirical evidence and theoretical considerations might force revisions to common-sense thinking about psychological knowledge. But the constraint of psychological knowledge does apply some pressure, because a theory is not at liberty to trample our common-sense conceptions without adequate motivation. In other words, there may be reasons to revise what we ordinarily think about psychological knowledge, but such reasons should be independent of the need to uphold any particular theory of the mind.

So far as knowledge of others is concerned, the constraint would seem to be as follows. In general, there is no serious doubt that other people do have thoughts and feelings just as we ourselves do (although we discuss the claims of *eliminativism* about the mental in chapter 2). And in particular cases we can know what it is that other people are thinking, whether they are happy or disappointed, what they intend, and what they are afraid of. Such knowledge is, however, not always easy to come by and in many instances behavioural or situational evidence may not be sufficient for any firm beliefs about another person's states of mind. Hence our psychological knowledge of others is not direct and immediate. It may or may not involve *conscious* inference about the thoughts and feelings of others. But even where no conscious inference is involved, our knowledge of other minds is dependent upon informational cues (from conduct, expression, tone of voice, and situation) – as can be seen from the fact that these cues can be manipulated by people who lie convincingly, pretend to be pleased when they are not, or make us forget for a while that they are just acting.

So far as knowledge of ourselves is concerned, while there can be such a thing as self-deception, we are vastly better informed than we are even about the psychological states of our nearest and dearest. In part this is because we have a huge store of past experiences, feelings and attitudes recorded in memory. But we would underestimate the asymmetry between self-knowledge and knowledge of others, if we represented it as just knowing *more*, in much the way that one knows more about one's hometown than other places. Self-knowledge differs from knowledge of others in that one seems to know in a different way and with a special sort of authority, at least in the case of one's *present* mental states. We seem to have a peculiarly direct sort of knowledge of what we are currently thinking and feeling. We do not seem to be reliant on anything in the way of evidence (as we would be if we were making inferences from our own situation and behaviour) and yet it hardly seems possible for us to be mistaken on such matters.

With the constraints of *naturalism* and *psychological knowledge* explained, we shall now review very briefly some of the main developments in twentieth-century philosophy of mind which form the back-drop to the main body of this book.

1.1 Dualism

Dualism comes in two forms – weak and strong. Strong dualism (often called 'Cartesian dualism') is the view that mind and body are quite distinct kinds of *thing* – while bodies are physical things, extended in space, which are subject to the laws of physics and chemistry, minds do not take up any space, are not composed of matter, and as such are not subject to physical laws. Weak dualism allows that the *subject* of both mental and physical *properties* may be a physical thing – a human being, in fact. But it claims that mental properties are *not* physical ones, and can vary independently of physical properties. Ever since Ryle's *The Concept of Mind* (1949) rejection of dualism has been the common ground from which philosophers of mind have started out. Almost everyone now agrees that there is no such thing as *mind-stuff*, and that the subject of mental properties and events is a physical thing. And almost everyone now maintains that mental properties *supervene on* physical ones, at least, in such a way that it is impossible for two individuals to share all of the same physical properties, but differ in their mental ones.

Much the most popular and influential objection to dualism (of either variety) concerns the *problem of causal interaction* between the mental and the physical. (Another objection is that dualism faces notorious problems in accounting for our psychological knowledge of others.) It seems uncontentious that there can be both physical causes which produce mental changes, and also mental events which cause bodily movements and, subsequently, changes in the physical environment. Perception illustrates the former causal direction: something happens and *you notice* it happening. Intentional action illustrates the mental-to-physical causal direction: after reflection you decide that the sofa would look better by the window, and this decision causes you to go in for some muscular exertions which in turn cause the sofa to get re-located. Such commonplaces are fundamental to our understanding of the relation between minds and their environment. But how such causal interactions could ever occur becomes mysterious on any consistently dualistic position, unless we are prepared to accept causal interaction between physical and mental events as a brute fact. And even if we *are* prepared to accept this, it is mysterious *where* in the brain mental events would be supposed to make an impact, given that enough is already known about the brain, and about the activities of nerve cells, to warrant us in believing that every brain-event will have a sufficient physical cause.

We cannot pause here to develop these and other arguments against dualism in anything like a convincing way. Our purpose has only been to give a reminder of why *physicalism* of one sort or another is now the default approach in the philosophy of mind. (Which is not to say, of course, that physicalism is unchallengeable. On the contrary, in chapter 9 we shall be considering arguments which have convinced many people that phenomenally conscious mental states – states with a distinctive subjective feel to them – are *not* physical.)

1.2 Logical behaviourism

The classic exposition of logical behaviourism is Ryle, 1949. His leading idea was that it is a mistake to treat talk about the mental as talk about inner causes and then go on to ask whether those causes are physical or not. To think this way, according to Ryle, is to commit a *category-mistake*. Talk about the mental is not talk about mysterious inner causes of behaviour, it is rather a way of talking about dispositions to behave and patterns of behaviour.

Behaviourism did have some attractions. It allowed humans to be included within the order of nature by avoiding postulation of anything 'ghostly' inside the organic machinery of the body. It also promised a complete (perhaps *too complete*) defence of our psychological knowledge of the minds of others, for knowing about others' minds was simply reduced to knowing about their behavioural dispositions. Furthermore, it seemed to be right, as Ryle pointed out, that people can correctly be described as knowing this or believing that, irrespective of what is going on inside them at the time – indeed, even when they are asleep.

The deficiencies of behaviourism were even more apparent, however. What always seemed most implausible about logical behaviourism was that *knowledge of one's own mind would consist in knowledge of one's behaviou'ral dispositions*, since this hardly left room for the idea of first-person authority about one's thoughts and feelings. The point that *some* of our mentalistic discourse is dispositional rather than episodic had to be conceded to Ryle. But then again, some of our mentalistic discourse is episodic rather than dispositional. Surely a sudden realisation, or a vivid recollection, or a momentary feeling of revulsion cannot be treated as a disposition. There are, it would seem, mental *events*. What is more, the fact that beliefs, knowledge and desires can be long-standing rather than fleeting and episodic is by no means a decisive argument that they are dispositions to behaviour. Their durational nature is equally compatible with their being underlying states with a lasting causal role or potential (as argued in Armstrong, 1973).

Logical behaviourism was offered as a piece of *conceptual analysis*. It was supposed to be an account of what had all along been the import of our psychological discourse. Allegedly, theoreticians had misconstrued our talk about the mind and loaded it with theoretical implications of unobserved mental mechanisms never intended in ordinary usage. That being the Rylean stance, the most serious technical criticism of logical behaviourism is that it fails on its own terms, as an exercise in analysis. According to behaviourism what look like imputations of internal mental events or states should actually be construed as 'iffy' or conditional statements about people's actual and possible behaviour. The first objection to the pretensions of behaviourist conceptual analysis, then, is that nobody has ever actually produced a single completed example of the behavioural content of such an analysis. In itself, this objection might not have been fatal. Ryle suggested such cases as *solubility* and *brittleness* as analogous to behavioural dispositions. To say that something is soluble or brittle is to say something about what it would do if immersed in water, or if struck by a solid object. Now, admittedly, there is a disanalogy, because there is just one standard way in which such dispositional properties as solubility and brittleness can be manifested (that is, by dissolving and by breaking into fragments). But no doubt there are more complex dispositional properties, both psychological and non-psychological. If there are various ways in which a complex dispositional property can be manifested, then spelling out in terms of conditionals what the attribution of such a dispositional property amounts to might well be an exceedingly difficult and lengthy task.

There is, however, a follow-up to the initial complaint about behaviourist analyses (and their non-appearance, in any detailed form), which not only blows away this flimsy line of defence, but also reveals a deeper flaw in behaviourism. Suppose I am walking along and come to believe that rain is about to start bucketing down. Do I make haste to take shelter? Well I may do so, of course, but that all depends. It depends upon such things as how much I care about getting wet, and also upon what I think and how much I care about other things which might be affected by an attempt to find shelter – such as my chances of catching the last train, or my reputation as a hard-as-nails triathlete. As Davidson (1970) pointed out, a particular belief or desire only issues in conduct in concert with, and under the influence of, other intentional states of the agent. There is no way, therefore, of saying what someone who holds a certain belief will do in a given situation, without also specifying what other beliefs and desires that agent holds. So analysis of a belief or a desire as a behavioural disposition requires invoking other beliefs and desires. This point has convinced practically everyone that Ryle was wrong. A belief or a desire does not just consist in a disposition to certain sorts of behaviour. On the contrary, our common-sense psychology construes these states as internal states of the agent which play a causal role in *producing* behaviour, as we shall go on to argue in chapter 2.

1.3 Identity theory

With dualism and logical behaviourism firmly rejected, attempts since the 1960s to give a philosophical account of the status of the mental have centred on some combination of *identity theory* and *functionalism*. Indeed, one could fairly say that the result of debates over the last forty years has been to establish some sort of functionalist account of mental concepts combined with token-identity theory (plus commitment to a thesis of supervenience of mental properties on physical ones) as the orthodox position in the philosophy of mind. There is quite a bit of jargon to be unpacked here, especially as labels like 'functionalism' and 'identity theory' are used in various disciplines for positions between which only tenuous connections hold. In the philosophy of mind, functionalism is a view about mentalistic concepts, namely that they represent mental states and events as differentiated by the functions, or causal roles, which they have, both in relation to behaviour and to other mental states and events; whereas identity theory is a thesis about what mental states or events *are*, namely that they are identical with states or events of the brain (or of the central nervous system).

There are two distinct versions of identity theory which have been the focus of philosophical debate – *type-identity* theory and *token-identity* theory. Both concentrate on an alleged identity between mental states and events, on the one hand, and brain states and processes, on the other, rather than between mind and brain *en masse*. Type-identity theory holds that each type of mental state is identical with some particular type of brain state – for example, that pain is the firing of C-fibres. Token-identity theory maintains that each particular mental state or event (a 'token' being a datable particular rather than a type – such as Gussie's twinge of toothache at 4 pm on Tuesday, rather than pain in general) is identical with some brain state or event, but allows that individual instances of the same mental type may be instances of different types of brain state or event.

Type-identity theory was first advocated as a hypothesis about correlations between sensations and brain processes which would be discovered by neuroscience (Place, 1956; Smart, 1959; Armstrong, 1968). Its proponents claimed that the identity of mental states with brain states was supported by correlations which were just starting to be established by neuroscience, and that this constituted a scientific discovery akin to other type-identities, such as *heat is molecular motion*, *lightning is electrical discharge*, and *water is H_2O*. In those early days, during the 1950s and 60s, the identity theory was advanced as a theory which was much the best bet about the future course of neuroscientific investigation.

Yet there were certainly objections which were troublesome for those who shared the naturalistic sympathies of the advocates of type-identity. A surprising, and surely unwelcome, consequence of the theory was an adverse prognosis for the prospects of work in artificial intelligence. For if a certain cognitive psychological state, say a thought *that P*, is actually to be identified with a certain human neurophysiological state, then the possibility of something non-human being in such a state is excluded. Nor did it seem right to make the acceptance of the major form of physicalist theory so dependent upon correlations which might be established in the future. Did that mean that if the correlations were not found one would be forced to accept either dualism or behaviourism?

But most important was the point that confidence in such type-correlations is misplaced. So far from this being a good bet about what neuroscience will reveal, it seems a very bad bet, both in relation to sensations and in relation to intentional states such as thoughts. For consider a sensation type, such as pain. It might be that whenever *humans* feel pain, there is always a certain neurophysiological process going on (for example, C-fibres firing). But creatures of many different Earthly species can feel pain. One can also imagine life-forms on different planets which feel pain, even though they are not closely similar in their physiology to any terrestrial species. So, quite likely, a given type of sensation is correlated with lots of different types of neurophysiological states. Much the same can be argued in the case of thoughts. Presumably it will be allowed that speakers of different natural languages can think thoughts of the same type, classified by content. Thus an English speaker can think that *a storm is coming*; but so, too, can a Bedouin who speaks no English. (And, quite possibly, so can a languageless creature such as a camel.) It hardly seems plausible that every thought with a given content is an instance of some particular type of neural state, especially as these thoughts would cause their thinkers to express them in quite different ways in different natural languages.

The only way in which a type-identity thesis could still be maintained, given the variety of ways in which creatures might have sensations of the same type and the variety of ways in which thinkers might have thoughts of the same type, would be to make sensations and intentional states identical, not with single types of neurophysiological state, but with some disjunctive list of state-types. So pain, for example, might be neuro-state H (in a human), or neuro-state R (in a rat), or neuro-state O (in an octopus), or . . . and so on. This disjunctive formulation is an unattractive complication for type-identity theory. Above all, it is objectionable that there should be no available principle which can be invoked to put a stop to such a disjunctive list and prevent it from having an indeterminate length.

The conclusion which has been drawn from these considerations is that type-identity theory is unsatisfactory, because it is founded on an assumption that there will be one–one correlations between mental state types and physical state types. But this assumption is not just a poor bet on the outcome of future research. There is something about our principles of classification for mental state types which makes it more seriously misguided, so that we are already in a position to anticipate that the correlations will not be one–one, but one–many – one mental state type will be correlated with *many different* physical state types. If we are to retain a basic commitment to naturalism, we will take mental states always to be realised in physical states of some type and so will conclude that mental state types are *multiply realised*. This is where functionalism comes in, offering a neat explanation of why it is that mental state types should be multiply realisable. Consequently, multiple realisability of the mental is standardly given as the reason for preferring a combination of functionalism and a *token*-identity thesis, according to which each token mental state or process is (is identical with) some physical state or process.

1.4 Functionalism

The guiding idea behind functionalism is that some concepts classify things by what they *do*. For example, transmitters transmit something, while aerials are objects positioned so as to receive air-borne signals. Indeed, practically all concepts for artefacts are functional in character. But so, too, are many concepts applied to living things. Thus, wings are limbs for flying with, eyes are light-sensitive organs for seeing with, and genes are biological structures which control development. So perhaps mental concepts are concepts of states or processes with a certain function. This idea has been rediscovered in Aristotle's writings (particularly in *De anima*). Its introduction into modern philosophy of mind is chiefly due to Putnam (1960, 1967; see also Lewis, 1966).

Functionalism has seemed to be the answer to several philosophical prayers. It accounts for the multiple realisability of mental states, the chief stumbling-block for an 'immodest' type-identity theory. And it also has obvious advantages over behaviourism, since it accords much better with ordinary intuitions about causal relations and psychological knowledge – it allows mental states to interact and influence each other, rather than being directly tied to behavioural dispositions; and it gives an account of our understanding of the meaning of mentalistic concepts which avoids objectionable dependence on introspection while at the same time unifying the treatment of first-person and third-person cases. Finally, it remains explicable that dualism should ever have seemed an option – although we conceptualise mental states in terms of causal roles, it can be a contingent matter what actually *occupies* those causal roles; and it was a conceptual possibility that the role-occupiers might have turned out to be composed of *mind-stuff*.

Multiple realisability is readily accounted for in the case of functional concepts. Since there may be more than one way in which a particular function, *ϕ-ing*, can be discharged, things of various different compositions can serve that function and hence qualify as *ϕ-ers*. Think of *valves*, for example, which are to be found inside both your heart and (say) your central heating system. So while mental *types* are individuated in terms of a certain sort of pattern of causes and effects, mental *tokens* (individual instantiations of those patterns) can be (can be identical to, or at least constituted by) instantiations of some physical type (such as C-fibre firing).

According to functionalism, *psychological knowledge* will always be of states with a certain role, characterised in terms of how they are produced and of their effects on both other such states and behaviour. Functionalism does not by itself explain the asymmetry between knowledge of self and knowledge of others. So it does need to be supplemented by some account of how it is that knowledge of one's own present mental states can be both peculiarly direct and peculiarly reliable. How best to deliver this account is certainly open to debate, but does not appear to be a completely intractable problem. (We view this problem as demanding a theory of consciousness, since the mental states one knows about in a peculiarly direct way are conscious ones – see chapter 9.) But if there is still unfinished business in the first-person case, one of functionalism's chief sources of appeal has been the plausible treatment it provides for psychological knowledge of others. Our attribution of mental states to others fits their situations and reactions and is justified as an inference to the best explanation of their behaviour. This view places our psychological knowledge of others on a par with theoretical knowledge, in two respects. Firstly, the functional roles assigned to various mental states depend upon systematic relations between such states and their characteristic causes and effects. So it seems that we have a common-sense theory of mind, or a 'folk psychology', which implicitly defines ordinary psychological concepts. Secondly, the application of that theory is justified in the way that theories usually are, namely by success in prediction and explanation.

We hasten to insert here an important distinction between the *justification* for our beliefs about the minds of others and *what causes* us to have such beliefs. In particular applications to individuals on specific occasions, we may draw inferences which are justified both by the evidence available and our general folk psychology, and may draw some such inferences (rather than others) *precisely because* we recognise them to be justified. But while our theory of mind can be justified by our predictive and explanatory successes in a vast number of such particular applications, we do not, in general, apply that theory because we have seen it to be justified. To echo Hume's remarks about induction, we say that this is not something which nature has left up to us. As we shall be arguing in chapters 3 and 4, it is part of our normal, native, cognitive endowment to apply such a theory of mind – in fact, we cannot help but think about each other in such terms.

So far we have been painting a rosy picture of functionalism. But, as usual, there have been objections. The two main problems with analytical functionalism (that is, functionalism as a thesis about the correct *analysis* of mental state concepts) are as follows:

(1) It is committed to the analytic/synthetic distinction, which many philosophers think (after Quine, 1951) to be unviable. And it is certainly hard to decide quite *which* truisms concerning the causal role of a mental state should count as analytic (true in virtue of meaning), rather than just obviously true. (Consider examples such as that *belief* is the sort of state which is apt to be induced through perceptual experience and liable to combine with *desire*; that *pain* is an experience frequently caused by bodily injury or organic malfunction, liable to cause characteristic behavioural manifestations such as groaning, wincing and screaming; and so on.)

(2) Another commonly voiced objection against functionalism is that it is incapable of capturing the felt nature of conscious experience (Block and Fodor, 1972; Nagel, 1974; Jackson, 1982, 1986). Objectors have urged that one could know everything about the functional role of a mental state and yet still have no inkling as to *what it is like to be in that state* – its so-called *quale*. Moreover, some mental states seem to be conceptualised purely in terms of feel; at any rate, with beliefs about causal role taking a secondary position. For example, it seems to be just the feel of pain which is essential to it (Kripke, 1972). We seem to be able to imagine pains which occupy some other causal role; and we can imagine states having the

causal role of pain which are not pains (which lack the appropriate kind of feel).

1.5 The theory-theory

In response to such difficulties, many have urged that a better variant of functionalism is *theory-theory* (Lewis, 1966, 1970, 1980; Churchland, 1981; Stich, 1983). According to this view, mental state concepts (like theoretical concepts in science) get their life and sense from their position in a substantive *theory* of the causal structure and functioning of the mind. And on this view, to know what a belief is (to grasp the concept of belief) is to know sufficiently much of the theory of mind within which that concept is embedded. All the benefits of analytic functionalism are preserved. But there need be no commitment to the viability of an analytic/synthetic distinction.

What of the point that some mental states can be conceptualised purely or primarily in terms of feel? A theory-theorist can allow that we have *recognitional capacities* for some of the theoretical entities characterised by the theory. (Compare the diagnostician who can recognise a cancer – immediately and without inference – in the blur of an X-ray photograph.) But it can be claimed that the concepts employed in such capacities are also partly characterised by their place in the theory – it is a *recognitional* application of a *theoretical* concept. Moreover, once someone possesses a recognitional concept, there can be nothing to stop them prising it apart from its surrounding beliefs and theories, to form a concept which is *barely* recognitional. Our hypothesis can be that this is what takes place when people say that it is conceptually possible that there should be pains with quite different causal roles.

While some or other version of theory-theory is now the dominant position in the philosophy of mind, this is not to say that there are no difficulties, and no dissenting voices. This is where we begin in chapter 2: we shall be considering different construals of the extent of our folk-psychological commitments, contrasting *realist* with *instrumentalist* accounts, and considering whether it is possible that our folk psychology might – as a substantive theory of the inner causes of behaviour – turn out to be a radically *false* theory, ripe for *elimination*. Then in chapter 4 we shall be considering a recent rival to theory-theory, the so-called *simulationist* account of our folk-psychological abilities. And in chapters 7 and 9 we consider the challenges posed for any naturalistic account of the mental (and for theory-theory in particular) by the intentionality (or 'aboutness') of our mental states, and by the phenomenal properties (or 'feel') of our experiences.

In fact one of the main messages of this book is that the theory-theory account of our common-sense psychology is a fruitful framework for considering the relations between folk and scientific psychologies, and so is to that extent, at least, a *progressive research programme* (in the sense of Lakatos, 1970).

2 Developments in psychology

We have to be severely selective in the issues in psychology which we examine in the following chapters. We have been mainly guided in our selection by two concerns: firstly, to examine aspects of psychology which might be taken as parts of the scientific backbone of the subject; and secondly, to address parts of psychology which are in a significant relation with common-sense psychological conceptions, either because they threaten to challenge them or because there is an issue about how well scientific psychology can be integrated with ordinary, pre-scientific thinking about the mind. Our general positions in relation to these two concerns are *realist* in regard to science and *Panglossian* on the relation between folk psychology and scientific psychology.

The term 'Panglossian' was coined by Stich (1983), recalling a character in Voltaire's novel *Candide* (called 'Dr Pangloss') who preached the doctrine that everything must in the end turn out for the best, since this world – having been created by a perfect God – is the best of all possible worlds. What Stich had in mind was that a modern Panglossian might *hope* that common-sense psychological conceptions would mesh quite well with what scientific psychology and cognitive science would reveal, but this was not much better than unfounded optimism in an easy and undisturbing outcome. However, we regard it as quite reasonable to hope for an integration of common-sense psychology and scientific psychology which will leave our pre-scientific psychological thinking substantially intact, although certainly enriched and revised. What chiefly supports the Panglossian prospect, in our view, is the fact that we are endowed with a highly successful theory of mind which has informative commitments to the causes underlying behaviour (a topic for chapter 2), and that this theory has developed as part of a modular capacity of the human mind which must be presumed to have been shaped by the evolutionary pressures bearing on our roles as interacting social agents and interpreters (themes for chapters 3 and 4). This falls short of a guarantee of the correctness of our native theory of mind, but it surely makes the Panglossian line worth pursuing.

We are also realists about the philosophy of science in general, and the philosophy of psychology in particular – which is not quite the same thing as being realist (in the way that we are) about *folk* psychology, since folk psychology is no science. What realists in the philosophy of science maintain is that it is the main task of scientific theories to provide a correct account of the nomological relations which genuinely exist between properties, and the causal powers of systems and entities, explaining these in terms of the generative mechanisms of the structures in virtue of which they have those powers. Anti-realists (such as van Fraassen, 1980) are apt to argue that no more can be asked of theories than that they should be empirically adequate, in the sense that they should be capable of predicting or accommodating all relevant observational data. The weakness of this anti-realist view is the assumption that there could possibly be a vantage point from which the totality of observational data is available. If it makes any sense at all to speak of such a totality, it is not something which is ever likely to be available to human investigators, who are continually finding novel ways of making relevant observations and devising new experimental techniques, without foreseeable limit. In fact, precisely one of the main advantages of realism is that it both allows and encourages an increase in the scope of observation.

Another major advantage of realism in the philosophy of science is that it gives a methodological bite to theorising, as Popper urged long ago (1956). If theories were merely instruments for prediction or the support of technology, then there would be no need to choose between different theories which served these purposes in equally good, or perhaps complementary, ways. But if we interpret theories as making claims about hidden or unobservable causal mechanisms, we will have to treat rival theories, not as different devices with their several pros and cons, but as mutually incompatible. This provides a spur to working out some way to decide between them – a spur to scientific progress, in fact. (See chapter 2 for more on different aspects of realism, and in particular for the case for realism about folk psychology.)

So much for our own general position. We now proceed to a swift survey of some very general trends in twentieth-century scientific psychology. Given the extent and range of recent scientific developments in this area, we must confine ourselves to some themes and topics which will recur in the following chapters. Some further areas of psychological research will then be surveyed, as appropriate, later in the book.

2.1 Freud and the folk

The theories of Sigmund Freud have attracted a degree of publicity which is out of all proportion to their actual influence within contemporary scientific psychology. In some respects Freud's theories have connections with themes of the present book which might have been worth pursuing.

For example, Freud clearly challenges some common-sense psychological conceptions. He is also clearly a realist both about intentional states and about his own theories. And he does make *use of* common-sense psychology, one of his major theoretical strategies being an attempt to extend ordinary styles of reason-explanation to novel applications – including behaviour previously considered to be unintentional, such as *Freudian slips*. It is also sometimes argued that some parts of Freud's theories have been absorbed by folk psychology, thus demonstrating that if folk psychology is a theory, it is not a completely fossilised or stagnating one. But this claim is questionable, since what folk psychology seems quite ready to acknowledge is the existence of unconscious beliefs and desires, rather than the distinctively Freudian idea of beliefs and desires which are *unconscious because repressed*.

The question of the methodological soundness of Freudian theory has been a matter of some controversy. Within philosophy of science it was given a special prominence by Popper (1957; 1976, ch.8), who treated Freud's theories (along with the theories of Marx and Adler) as a prime example of how theorising could go wrong by failing to satisfy the famous *Demarcation Criterion*. Genuinely scientific theories such as Einstein's theory of relativity were, according to Popper, distinguished by their falsifiability; that is, by there being tests which, if carried out, might possibly give results inconsistent with what such theories predicted, thereby refuting them. If theories could not be subjected to test in this way, then they were merely *pseudoscientific*. Popper's philosophy of science is now generally regarded as inadequate, because it fails to do justice to the role of auxiliary hypotheses and the long-term appraisal of research programmes. So the Popperian critique no longer seems so damaging. (Though see Cioffi, 1970, for an account of Freud's own defence of his theory of the neuroses which undeniably makes it appear worryingly pseudoscientific.)

We will not be engaging with Freud's ideas, however, or any issues concerning psychoanalysis in this book. Where Freudian theories do have any testable consequences they have consistently failed to be confirmed, and the overall degeneration of the Freudian programme has reached a point at which it is no longer taken seriously by psychologists who are engaged in fundamental psychological research. The tenacity with which these theories survive in areas of psychotherapy (and also in literary theory and other areas of the humanities), in increasing isolation from any research which might either justify their application or testify to their clinical effectiveness, is a matter of some concern. But we do not propose to go into this in the present work. (For discussion of the methodology and clinical effectiveness of psychoanalysis, consult Grünbaum, 1984, 1996; Erwin, 1996.)

2.2 *Methodological behaviourism*

We have already mentioned the arguments against *behaviourism in philosophy* (logical behaviourism). But there is also a behaviourist position in psychology. Indeed, for much of the twentieth century – under the influence of such theorists as Watson, Guthrie, Hull, Skinner, and Tolman – this was the dominant position in psychology, and it remains influential in studies of animal behaviour.

Although some theorists undoubtedly subscribed to both brands of behaviourism – methodological *and* logical – the two positions are distinguishable. A modest form of methodological behaviourism is not vulnerable to the arguments which sank logical behaviourism in philosophy. Methodological behaviourism need not deny that there are mental states and internal psychological mechanisms, it just declines to delve into what they might be – on the grounds that, being unobservable, they are not amenable to controlled scientific investigation. It proposes to treat the central nervous system as a 'black box', the contents of which are hidden from scrutiny. Rather than indulge in mere speculation about what goes on inside there, better to concentrate on what can be quantitatively measured and objectively analysed – the behaviour emitted by the organism in response to various stimuli. Stimuli and responses are undoubtedly observable, and stimuli can be controlled and varied to determine corresponding variations in response. So laws governing associations between stimuli and responses should make a respectable subject for empirical science.

We reject methodological behaviourism on two main grounds. Firstly, in terms of the philosophy of science it is a typically positivistic, anti-realist stance, confining the aims of inquiry to lawlike generalisations concerning what is – on a narrow view – taken to be observable. This we regard as unwarranted pessimism about the growth of scientific knowledge. Often scientific theory has been at its most progressive precisely when postulating previously unobserved entities and mechanisms. A self-denying programme which restricts us to studying associations between stimuli and responses is, in the long term, only an obstacle to progress. Secondly, there is a problem relating to psychological theory, and particularly to learning and cognitive development. Treating the central nervous system as a black box puts investigators seriously at risk of neglecting the extent to which cognitive functions and developmental profiles depend upon the internal structure of a complex system which is the product of evolutionary design. In so far as behaviourism neglects this structure by adopting an empiricist, associationist view of learning, we can leave the evidence against it to be presented in chapter 3, where we make out the case for the principles of *modularity* and *nativism*. The message, in brief, is that a significant part of our psychological capacities *mature without learning*.

Behaviourism would never have achieved the influence it did without having some paradigmatic experimental achievements to display, of course. Examples of *Pavlovian* or *classical conditioning* are well known: an animal responds to an *unconditioned stimulus* (such as the sight of food) with an *unconditioned response* (such as salivating); it is then trained to associate a *conditioned stimulus* – some other, initially neutral stimulus (such as a bell ringing) – with the unconditioned stimulus (sight of food); until eventually the conditioned stimulus (the bell) produces a *conditioned response* (such as salivating – though conditioned responses need not be identical with unconditioned responses). Behaviourists could also point to replicable instances of *Thorndikian* or *instrumental learning* in support of their research strategy. In one of the earliest of these experiments (Thorndike, 1898), hungry cats were placed inside a box with a grille on one side which afforded a view of some food. A door in the grille could be opened by pulling on a looped string within the box – a trick which the cat has to learn in order to get the food. On repeated trials, Thorndike found that cats did learn this trick, but on a trial-and-error basis and only gradually, with the number of fruitless attempts to get at the food steadily decreasing.

Such results prompted Thorndike to formulate the *law of effect*, according to which responses become more likely to recur if followed by a rewarding outcome, less likely if followed by no reward or discomfort. This law, in various formulations (such as Hull's *law of primary reinforcement* or Skinner's *principle of operant conditioning*), is the basic idea behind behaviourist learning theory. But although it certainly lent itself to attempts at experimental demonstration and quantitative measurement, behaviourist learning theory exhibited little in the way of genuine theoretical progress. It remained unclear how instrumental learning could be transferred, from methods of training animals to perform somewhat unnatural tricks in the laboratory, to yield an understanding of what controlled behaviour in natural environments. Above all, much of behaviour (human or non-human) seemed just too complex to be regarded as *a response*, or even a series of responses. Even a one-time behaviourist like Lashley questioned behaviourism's capacity to give an account of behaviour involving complex serial order, such as piano-playing (Lashley, 1951).

A very important kind of behaviour in which complex serial order is salient, of course, is linguistic behaviour. Chomsky's hostile review (1959) of Skinner's *Verbal Behaviour* (1957) was extremely influential. For it revealed just how inadequate are methodological behaviourism, and its learning-by-reinforcement, to the task of giving any account of the actual and potential verbal behaviour of an ordinary native speaker. On any view, it seemed clear that linguistic production and linguistic comprehension requires the presence of a rich knowledge-base in the ordinary human speaker.

Convinced of the degenerating trend of the behaviourist research programme, theorists increasingly turned towards hypotheses about what cognitive systems were at work inside the ‘black box’. They have been rewarded by the sort of *expansion of evidence* about internal structure which, as we mentioned above, is one of the advantages of a realist approach to scientific investigation. Evidence concerning psychological mechanisms has now come to encompass such diverse sources as: developmental studies; population studies and their statistical analysis; the data concerning cognitive dissociations in brain-damaged patients; data from neural imaging; and many different sorts of experiments designed to test hypotheses about internal processing structures, by analysing effects on dependent variables. Examples of each of these sorts of evidence will be found in the chapters which follow (particularly in chapters 3–5).

2.3 The cognitive paradigm and functional analysis

The broad movement which superseded behaviourism, and which has, to date, proved far more theoretically progressive, is *cognitivism*. Cognitive psychology treats human brains and the brains of other intelligent organisms – as, at bottom, information-processing systems. It must be admitted that the emphasis on cognition in modern psychology has tended by comparison to leave aspects of psychology in the category of *desire* somewhat in the shade. We do actually offer a tentative suggestion as to how desire, conceptualised according to folk-psychological theory, may fit in with a modular cognitive architecture in chapter 3 (section 5.3). Whether this integrative effort is supported by future research remains to be seen. What is clear is that discoveries in cognitive psychology already constitute a fundamental part of scientific psychology, and will surely continue to do so in the future.

Yet again the word ‘function’ appears, though functional analysis in cognitive psychology is not the same thing as functionalism in the philosophy of mind. In cognitive psychology the object of the exercise is to map the functional organisation of cognition into its various systems – such as perception, memory, practical reasoning, motor control, and so on – and then to decompose information-processing within those systems into further, component tasks. Functional analysis of this sort is often represented by means of a ‘boxological’ diagram, or flow-chart, in which the various systems or sub-systems are shown as boxes, with arrows from box to box depicting the flow of information. We produce, or reproduce, a few such diagrams in this book (see figures 3.3, 4.1, 9.3 and 9.4). It might be complained of this style of boxological representation that, if not completely black, these are at least *dark* boxes within the overall container of the mind, in that we may not know much about how *their* innards work. This is true – but it is no objection to the project of functional analysis that there is still plenty more work to be done! Dennett (1978f) has likened this style of functional analysis to placing lots of little homunculi in the cognitive system, and then even more 'stupid' homunculi within the homunculi, and so on. The ultimate objective of the analysis is to decompose the processing into completely trivial tasks.

It is tempting to suppose that it was the advent of the computer which made modern cognitive psychology possible. This might be offered as some excuse for the limitations of behaviourism, in so far as this essential tool for investigating what intervenes between stimulus and response was not available until the later decades of the century. But despite the invaluable aid supplied by computer modelling, this is at best a half-truth. Thus Miller, in one of the most influential papers in cognitive psychology (1956), proposed the thesis that there is a severe restriction on human information processing, in that about seven or so items of information (7 ± 2) are the maximum that we can handle either in short-term recall or simultaneous perceptual judgements. Computer modelling would be of little help in establishing this feature of human information processing (which had, indeed, been partially anticipated by Wundt – 1912, ch.1). There have been many other test results which vindicate the cognitivist approach by relating human performance to an assessment of the processing task involved; for example, relating the transformations involved in production or comprehension of speech, according to grammatical theory, to the ease, accuracy, or speed with which subjects perform (see Bever, 1988, for references to several such studies).

So psychology has taken a cognitive turn, and there is very general agreement that it was a turn for the better. The result has led to fertile interconnections between cognitive psychology itself, research in computer science and artificial intelligence, neurophysiology, developmental psychology (as evidenced in relation to mind-reading in chapter 4), and evolutionary psychology (see chapter 5 for the example of *cheater-detection*). But within cognitivism there is a dispute between so-called *classical* and *connectionist* cognitive architectures.

2.4 Cognition as computation

According to the classical, or symbol-manipulation, view of cognition, the mind *is* a computer – or better (to do justice to modularity: see chapter 3), a system of inter-linked computers. Apart from the availability of computers as devices for modelling natural cognition and as an analogy for information-processing in the wild, there are a number of general considerations in favour of supposing that the mind processes information by operating on symbolic representations according to processing rules, in much the way that computers do when running programmes.

One sort of consideration concerns the processing task which perceptual systems must somehow accomplish. The role of these systems in cognition is to provide us with information about the environment. But the actual input they receive is information which derives immediately from changes in the transducers in our sensory organs. They must, therefore, somehow recover information about the environmental causes of these changes. How is that to be done? One answer which has been pursued within the cognitive paradigm is that these systems work by generating hypotheses about external causes of internal representations. Cognitive science can investigate this processing by first providing a functional decomposition of the processing task, and then working out algorithms which would yield the desired output. Perhaps this consideration in favour of the computational view is no longer as compelling as it once seemed. *We* could not think of any other way in which the processing task could be accomplished, but perhaps Mother Nature could. What is more, there is now a known (or so it seems) alternative to rule-governed manipulation of internal representations in the form of connectionist networks. But even if information processing does not *have* to be done by means of symbol manipulation, the theory that it does operate in this way can claim such a considerable degree of empirical success in modelling perception and cognition that nobody would lightly abandon it (see, for example: Newell and Simon, 1972; Simon, 1979, 1989; Marr, 1982; Newell, 1990).

Another consideration in favour of a computational approach to cognition derives from Chomsky's seminal part in the cognitivist revolution. Chomsky maintains that both production and comprehension of utterances (linguistic *performance*) depend upon the speaker's – and hearer's – *competence*; and that this competence consists in a tacit knowledge of the grammatical principles of the speaker's native language. So Chomsky is committed to linguistic processing on internal representations which is governed by these grammatical principles. And, as mentioned above, a body of empirical evidence does appear to show that Chomsky is right, by attesting to the psychological reality of this sort of processing (Bever, 1988; Bever and McElree, 1988; MacDonald, 1989).

Much the most vociferous advocate of classical computationalism, however, has been Fodor, who has consistently argued, not only that cognition consists in computation over symbolic representations, but also that it requires an innate symbolic medium or *language of thought* (generally referred to as 'LoT', or 'Mentalese'). One of his early arguments for Mentalese was that it is required for the acquisition of any new word in a natural language, since in order to grasp a term one has to understand what it applies to, and one can only do that by means of a hypothesis which expresses an equivalence between the newly acquired term and a concept in some other medium – a medium which must precede acquisition of natural language concepts (Fodor, 1975). Few have found this particular argument convincing. But the conclusion might be true, for all that. Fodor has since offered arguments for computationalism combined with Mentalese which draw on quite general, and apparently *combinatorial*, features of thought and inference (Fodor, 1987; Fodor and Pylyshyn, 1988). In chapter 8 we will be considering the case for a language of thought and also exploring the extent to which natural language representations might be capable of serving some of the functions which computationalists have assigned to Mentalese.

In chapter 8 we also debate whether connectionism should be taken as a serious – or, as some maintain, superior – rival to the computational model of mind. Here we limit ourselves to some introductory remarks on how connectionism differs from the classical computational approach.

2.5 Connectionism and neural networks

One sometimes hears it objected, against the computational view, that brains do not look much like computers. This is a rather naive objection. There is no reason to expect computers fashioned by nature to be built of the same materials or to resemble in any superficial way the computers made by human beings. However, it is undeniably true that at the level of neurons, and their axons and dendrites, the structure of the brain does resemble a network with nodes and interconnections.

As early as the 1940s and 1950s the perceived similarity of the brain to a network inspired a few researchers to develop information-processing networks especially for the purposes of pattern recognition (McCulloch and Pitts, 1943; Pitts and McCulloch, 1947; Rosenblatt, 1958, 1962; Selfridge and Neisser, 1960). However, for some years work on processing networks was sidelined, partly by the success of the classical computational paradigm and partly by limitations of the early network models (as revealed in Minsky and Papert, 1969).

These limitations have since been overcome, and in the wake of Rumelhart and McClelland's work on parallel distributed processing (1986) there has been an upsurge of interest in connectionist modelling. The limitations of the early network models resulted mainly from their having only two