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 home-town 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