Introduction

This unit examines one of the most fascinating areas of philosophy: the nature of the mind. We ask what the mind actually is and what its relationship is to the physical world. The puzzles that arise are mainly related to the fact that the mind is so different from everything else in nature. It appears not to be capable of being directly studied by science—or at least by the physical sciences. Particularly baffling is the question of the mind's relationship to the body. Is your mind just your brain or is it something over and above that?

Some of the texts in unit 4 are highly relevant to these topics. If you are studying either Hume's *Enquiry concerning Human Understanding* or Descartes' *Meditations*, you will find discussions of several of the themes explored here.

Substance dualism

1. Descartes’ theory of substance dualism

*Dualism* means a theory that recognises two sorts of basic things. In the philosophy of mind, it refers to a theory that divides the world into two entirely different realms: the mental and the physical. *Substance dualism* says that mind and body are two distinct and separate entities. This theory, developed by the French philosopher Rene Descartes (1596-1650), is also referred to as *Cartesian dualism*. According to Descartes, mind and body are distinguished by the fact that the former is *simple* and *indivisible*, whereas the latter (in fact matter in general) is *spatially extended*. A person's mind interacts with her body in two ways: it is affected by her body when she receives information through any of her senses and it affects the body in its turn when she chooses to move any part of her body. Descartes thought that the mind and body interacted at a specific place—the pineal gland (located in the centre of the brain, but now known to be the source of the hormone melatonin, which affects sleep patterns). Of course, substance dualism can be discussed without reference to the pineal gland or any other specific bodily organ. (Incidentally, Descartes thought non-human animals do not have minds, despite the fact that, like humans, they have pineal glands.)

Clearly, Descartes' view of the mind was close to what religious people would call the *soul*, an immaterial thing that we each have and which constitutes the essence of what we are.

*Reasons for adopting substance dualism:*

- It provides a reasonably intuitive account of the mind-body relation. Most of us probably do think we are minds 'inside' bodies (though is this perhaps because we have had too much exposure to that view and not enough to the alternatives?).
- It seems impossible to understand how the mind could be anything other than completely distinct from the body. Mental states simply do not have the same sorts of properties as physical things (size, weight, thickness etc.), while physical things are in their turn utterly different from mental states (they do not 'feel' like anything; you cannot coherently talk about what it would be like to be a table or a proton).
Problems with substance dualism:

If the mental and the physical are so utterly different, how is it even possible for them to interact? Thus suppose we adopt Descartes' conception of the fundamental difference between the two, such that the mind is essentially simple and indivisible, while the body is spatially extended. We can understand how two spatially extended things can interact: we are familiar with this from our experience of the everyday world. But how could something that is spatial interact with something that isn't? The lack of spatial properties in the case of the mind would seem to make it impossible to explain how the interaction takes place. Some might argue that the nature of the causal interaction could be explained without reference to spatial extension as long as there was some sort of 'interface' between the two elements of the causal relation. But how could there be an interface when it seems that mind and body have nothing in common? We appear to be left with a profound and unsatisfactory mystery.

The problem of solipsism. This is the theory that I alone exist (or in your case, that you alone exist). This is not a popular theory (!) but it can seem that if we start with something like substance dualism we are inevitably led to it. For after all, whereas the existence of my mind is immediately evident to me (Descartes himself famously said 'I think, therefore I am'), the existence of other things, even my own body, might be a delusion. (Descartes expressed the worry by imagining that he might always be dreaming when he thought he was awake.) Equally worrying is the thought that you and everyone else might be just figments of my imagination. Some people react to such thoughts by dismissing them as a kind of insanity. But in philosophy, as Descartes emphasized, we are not entitled to do this: every possibility that we (or rather I!) can conceive must be considered without bias or prejudice of any kind.

Responses to these problems:

In response to the interaction problem, instead of maintaining that the mental and the physical literally interact, we might adopt a different conception of the relationship between mind and body. (Of course, although this involves retaining substance dualism, it does not retain Descartes' precise version of it.) Alternatives include:

1. **Psychophysical parallelism.** According to this view, the mental world and the physical world are like two clocks, wound up by God so that they always show exactly the same time. In other words, mental states and physical states always correspond with one another, but without there being any real interactions between them.

2. **Occasionalism.** On this view, every time a mental event seems to cause a physical event or vice-versa, what really happens is that God intervenes to bring about the necessary mental or physical 'effect'. Again, there is no actual causal interaction between the mental and the physical.

3. **Epiphenomenalism.** According to this view of the mind-body relationship, although physical events can cause mental events, mental events do not have the ability to cause physical events. If I move my arm, the real cause of the
movement is an event in my brain (which is caused in its turn by an earlier physical event and so on). My mental decision to move my arm is just an 'epiphenomenon' with no effects of its own. In recent times, some evidence for this has emerged from experiments which seem to show that the physical processes initiating a voluntary movement actually precede the psychological feeling of deciding to move, albeit by only a few milliseconds. Epiphenomenalism, however, only removes the interaction problem in one direction—it does not require the mental to affect the physical, but physical to mental causation may still be considered unexplained. And in any case, to lose mental to physical causation may itself seem quite unacceptable. For we feel sure that our thoughts really do make a difference to what our bodies do. To try to explain this away as an illusion would provoke incredulity in most people.

In response to the problems of solipsism and other minds, the substance dualist might say that although we cannot prove the existence of physical objects and of other minds, it is reasonable to infer their existence from the evidence provided by our own mental states. We can perhaps use the method of inference to the best explanation to argue that physical bodies probably exist. (That is to say, the best explanation of why bodies appear to exist is that they actually do exist.) Perhaps we can also argue, by a form of analogy, that since our own mental states are correlated with the states of our own body, then it is likely that other human bodies have mental states correlated with them.

2. Could there be a private language?

During the twentieth century, ideas emerged which seemed to cast grave doubts on the suggestions of the above paragraph. These ideas came to a large extent from the Austrian-British philosopher Ludwig Wittgenstein (1889-1951) and persuaded many that Cartesian dualism was deeply incoherent and needed to be abandoned. We will now look at these important ideas in some detail.

They revolve around the suggestion that it is wrong-headed to think that we learn first how to ascribe mental states to ourselves and then, by some form of analogy, learn to apply them to others. If we think about how young children actually learn the use of mental predicates (such as 'has a pain' or 'is happy'), we can see that they learn how to apply them to themselves ('I have a pain', 'I am happy' etc.) in tandem with the process of learning how to apply them to others ('Mummy has a pain', 'Mummy is happy').

Cartesian dualists might reply that although this may be so, it is still true that our own mental states are the only ones that we directly know about and we need something to justify our belief in the mental states of others, however natural—or even inevitable—such a belief might be.

Wittgenstein had an important argument against this view, known as the private language argument. He imagined a person having a private language that he alone could understand. By definition, he could not use this language to communicate with others, but the idea is that he
could read and write in such a language for his own benefit. The words in this person's language refer to properties of his sensations, e.g. 'blue', 'bright', 'loud', 'soft'. Whenever he has certain kinds of sensations, he records them in a diary. This might be a useful thing to do. For example, he might find that whenever he has a certain sensation, say of blue, a manometer shows that his blood pressure rises. Then he could learn to recognise a rise in his blood pressure without the need for any apparatus. Now Wittgenstein makes a very interesting observation about this situation. He points out that it doesn't matter whether the hypothetical private language user is using his sensation-words consistently or not. Things would work just as well if he was using the word 'blue' sometimes for blue sensations and at other times for red ones, as long as his blood pressure rises whenever the sensation that he gets seems to him to be blue. But, Wittgenstein then argues, this implies that 'the hypothesis that I make a mistake is mere show.' (Wittgenstein, *Philosophical Investigations*, p. 270.) For this 'language user', there is no distinction between a correct and an incorrect use of the word 'blue'. But in that case, he is not using the word to refer to anything. The same argument applies to every word in his supposed private language. And that shows that the very idea of such a language is nothing but a fantasy.

So, the argument goes, our own mental states are not 'private' objects that can only be known by us and described by us in a 'private language' that is inaccessible to others. Someone else can know that I am in pain just as well as I can (and vice-versa). There is thus no genuine problem of other minds. Solipsism isn't even intelligible and Descartes' picture of two separate substances, mind and body, nonsensical. As I said, this view has been very influential, but not everyone accepts it. It is hard to believe that a private language is not possible, at least in principle. But can Wittgenstein’s argument be found faulty? I’ll leave you to decide.

**Materialism: the options**

If we reject substance dualism, as most philosophers these days do, what should we put in its place? The most popular answer—amongst professional philosophers writing in English at least—is that we should adopt some form of materialism, in which only physical entities are regarded as real. Materialism is often divided into two main kinds: reductive and non-reductive. Unfortunately there is no agreed definition of the difference. But here is one way of looking at the matter. A reductive materialist will say that everything that exists, including the mind, is physical. In addition, she will claim that everything that we need to say about the world can be expressed in purely physical terms. A non-reductive materialist, in contrast, will adopt one of two positions, both of which involve agreeing that everything is physical. The first position is that in talking about the mind we are adopting a way of talking that is indispensable—it cannot be reduced to purely physical talk, despite the fact that everything ultimately is physical. The second position is radically different. It involves denying that the mind even exists. All of these possibilities will be looked at in due course.
Reductive materialism

1. Logical behaviourism

Reductive theories come in three main forms:

- Logical behaviourism
- Basic identity theories
- Functionalist theories

We will look at each of these in turn.

Logical behaviourism is the theory that to talk about the mind and mental states is simply to talk about observable human behaviour. It is chiefly associated with British philosopher Gilbert Ryle (1900-1976) and his classic book *The Concept of Mind*. This is a lively attack on Cartesian dualism and an attempt to supplant it with a less problematic view.

Ryle views the mind as being nothing over and above the body. It is not (in his memorable phrase) a 'ghost in the machine': to talk about the mind is just to talk in a special way about the body or about observable human behaviour. Ryle uses a very helpful analogy. Suppose someone is being shown around a university. He is shown the various colleges, the administrative buildings, the place where exams are held, the student union and so on. He is shown students and lecturers going about their business. When the tour finishes he says, 'Yes, that's all very interesting, but where is the university?' Clearly this person is suffering from a deep misunderstanding. He has already been shown the university, although he doesn't realise it. The university is nothing over and above the buildings, institutions and activities that compose it. In the same way, Ryle goes on to argue, a mind is not anything *in addition* to the body and its activities. To talk about a person's mind or her mental states is just to talk in a special way about her behaviour. For example, to say that she is *courageous* is to say that she is disposed to act bravely in circumstances that require it. To say that she is *kind* is to say that she is disposed to act kindly when the occasion demands. To say that she *knows* how to play chess is to say that, when required or when she wishes to, she will play a game of chess correctly according to the rules. In other words, the mental terms 'courageous,' 'kind' and 'knows' can be reduced to statements about how a person acts or behaves under certain conditions. So it is, Ryle claims, in relation to all mental terms.

How convincing is the logical behaviourist account? It certainly seems to work well for psychological dispositions such as those mentioned above (courage, kindness and so on). But it is not entirely clear that Ryle's theory can account for episodic mental states such as catching a glance of something or feeling a stabbing sensation. Ryle is fully aware that Cartesian dualists are likely to use these sorts of cases to try and defend their view. His response is a little surprising. He says: 'To use an objectionable phrase, there is nothing “mental” about sensations.' (Ryle, *The Concept of Mind*, p.195.) In other words, sensations are not really a part of the mind. It seems that, according to Ryle, insofar as we should be using the word 'mental' at all, we ought to confine it to descriptions of people's general dispositions and character. He also warns against treating a phrase like 'my twinge' as standing for any sort of *thing*. We cannot meaningfully talk
of observing or inspecting a twinge or any other sensation. To say that 'I had a twinge' does not assert a relation between myself and an object of some kind in the way that 'I had a hat' does. According to Ryle, it is perfectly proper to talk about having sensations without supposing that this is to apprehend or sense objects of some mysterious sort.

I will leave it to you to decide whether this is a satisfactory way of responding to the dualist's likely insistence that our experiences of having sensations prove logical behaviourism to be untenable.

2. Basic identity theories

Identity theories assert, roughly speaking, that the mind is identical to the brain. There are two main kinds: type identity theories and token identity theories. The former say that each type of mental state is identical to a type of physical state. The latter say that each individual mental state is identical to an individual physical state, but identities do not necessarily hold at the level of types. We shall be concerned here with type identity theories.

In a way, our third category of theory--functionalism--is a kind of identity theory, as we shall see. I refer to the type of theory discussed in this section as 'basic' identity theories in order to distinguish them from functionalism.

Basic identity theories assert that particular types of mental states (pains, colour sensations, types of tactile feeling and so) are identical to particular types of physical states in the brain. For example, suppose a physiologist discovers that a subject reports being in pain when and only when particular fibres in her body or brain called 'C-fibres' are being stimulated. Whereas a substance dualist would say that the two types of state (pains and stimulations of C-fibres) are distinct and merely correlated with one another, the type identity theorist will say that pain simply is stimulation of C-fibres. Of course, people who are in pain do not generally know that their C-fibres are being stimulated. However, this only reflects the fact that the identity of pain and C-fibre stimulation is an empirical discovery rather than a self-evident truth. A good analogy might be with the nature of lightning or thunder. Empirical science shows that lightning and thunder are identical to electrical discharges taking place in the atmosphere. The fact that most people do not know this does not prevent it from being true.

The American philosopher Saul Kripke (b. 1940) has an interesting argument against this view. He maintains that the analogy with the identity between lightning and an electrical discharge does not work. This is because, whereas in the latter case we can clearly distinguish between what the lightning seems to be (a flash in the sky) and what it actually is (an electrical discharge), we cannot similarly distinguish between what the pain seems to be (an unpleasant sensation) and what it actually is. Why not? Because you can't intelligibly distinguish between the way any sensation seems to be and the way it is. Pain, being a subjective experience, just is a type of seeming! Thus Kripke concludes that the theory of the identity between types of mental states and types of physical states is misguided.
Here is another objection to basic type identity theories. Suppose you came across someone who looked and behaved exactly like a normal human being, but it turned out that wherever in our bodies we have carbon, this being had silicon. Would you stop thinking that this person was conscious, that she had a mind, just because the material of which her body was made was not the same as ours? This seems unlikely, if, as we are supposing, the person's appearance and behaviour are in all respects human. This appears to show that any identification of types of mental states with types of physical states involving reference to particular kinds of material will not work, for it will be too restrictive in the range of beings that it counts as having conscious mental states.

3. Functionalist theories

Functionalist theories were specifically developed to overcome the above objection concerning the possibility of conscious states being realised in less familiar physical forms, such as silicon. Instead of identifying a given type of mental state with a given type of constitutionally specific physical state, a functionalist identifies it with a certain type of functional state. What does this mean?

American philosopher Hilary Putnam (b. 1926) points out that a pain is typically associated with damage to a body part and involves an inclination to withdraw that body part and to avoid what is thought to have caused the damage. He suggests that we therefore identify pains of the relevant sort with the sorts of states that cause the organism to have such inclinations. This is a functional identification because, instead of saying that a pain is a certain kind of physical state, involving certain specific kinds of physical or chemical processes, it says that it is a certain kind of functional state, involving inclinations towards certain kinds of behaviour. As such, it could be realised in a large and indeed indefinite range of physical ways. Similarly, desires could be identified with certain states that have the functional characteristic of tending to cause certain kinds of actions, namely those that the organism believes will secure its goals. (These are only rough characterisations, by the way.) Now there is no longer any problem about recognising that the silicon-based organism behaving exactly like a normal human being is conscious, for this organism will have the sorts of functional states that the present theory treats as equivalent to mental states such as beliefs, desires, pleasures and pains.

Note that functionalists may still be thought of as identifying mental states with physical states of some sort, provided that it is understood that these physical states are conceived of rather abstractly and not in terms of specific physical or chemical materials and processes.

The functionalist theory may remind you of logical behaviourism in the way that it invokes inclinations to behave in certain ways. But it is not the same as logical behaviourism. The latter sees consciousness purely in terms of dispositions to behave in certain ways and has little or nothing to say about inner states. Functionalism, in contrast, has no problem with inner states. It treats them as being identical to the appropriate functional states. The latter may be considered independently of the behaviour that they give rise to and thus do seem to be suitable candidates for the conscious states of which our mental lives appear to be composed.
4. Problems for reductionism

Any convincing philosophy of mind needs to say something about qualia (singular = quale). Qualia are the subjective, 'felt' qualities of experiences. Consider, for example, the experience of tasting chocolate. The quale in this experience is the chocolaty taste itself. Notice how indefinable this is. You would not be able to explain exactly what this taste was like to someone who had not experienced it herself. Another example: the experience of looking at green grass. The quale here is the greenness and it too is indefinable. It seems that science cannot wholly explain qualia. This can be shown by asking: why does the grass appear like that and not some other way--red, say? Science seems unable to explain this. It can tell us why our brains are affected in a certain way when our eyes receive light of a certain frequency, but this leaves out the greenness, which is the very thing that needs explaining.

Another aspect of the mind that seems to resist reduction to the concepts of science is intentionality. This refers to the property that all conscious states seem to have of being about something. This is most obvious in the case of thoughts. By its very nature a thought is always about something or someone. But it is arguably also true of other mental states. For example, the state of being in pain always seems to involve consciousness of or about something, most often a body part. Now this feature of 'aboutness' or intentionality is something that is entirely absent from scientific accounts of brain states, neural firings and so on. So it is quite plausible to suggest that intentionality is another feature of consciousness that prevents any reduction of it to the physical.

The so-called 'hard problem of consciousness', to which consideration of such things as qualia and intentionality naturally leads, is this: how is it that some physical organisms are subjects of experience? Note that functionalism is of limited use here. It tells us nothing about the subjective, 'felt' qualities of experience. Though it might be thought to accommodate intentionality, it does not even begin to address the issue of qualia.

At this point, it is necessary to introduce the knowledge argument. This is an argument which has been used to try to show that all forms of physicalist reduction of the mind are doomed to fail. It runs as follows. Consider a hypothetical person called Mary. Mary is a neurophysiologist, an expert in the workings of the human brain. Mary also happens to have been blind from birth. Perhaps it is as a result of her disability that Mary is particularly interested in the neurophysiology of sight, especially the perception of colour. In fact, she knows all there is to know about what happens in the brain when a person sees, say, the colour green. Now the question is this: does Mary know everything there is to know about the experience of seeing green? And the intuitive answer that most people give is that she certainly does not know everything about it, because she does not know what it is like to see green. She does not know the experience, as it were, 'from the inside'. All the knowledge that she has about the neurophysiology of sight, however detailed it may be, cannot give her that. It is worth thinking about why this is. It relates to the nature of scientific knowledge, which is purely structural. In other words, it concerns the relationships between entities whose intrinsic nature is not known. (No-one knows the intrinsic nature of an electron or a quark. We know only what these things will do when they come into contact with other entities of various kinds.)
Since Mary's detailed neurological knowledge concerning the phenomenon of sight misses out something crucial—what it is like to see green or any other colour—it does not include all the essential facts about seeing. And since Mary's blindness from birth prevents her from filling this gap using her imagination, it follows that Mary cannot know all there is to know about seeing colours. However, through her detailed neurophysiological knowledge, she does know all there is to know in physical terms about seeing and so it follows that the physical aspect of seeing cannot wholly define what the experience of seeing actually is. The argument can be repeated for any mental state that involves qualia. This looks like a very strong argument against any form of physicalist reductionism.

We should also consider zombies. In philosophy (as opposed to 'Shaun of the Dead') a zombie is a creature that gives every appearance of being a normal person, but actually has no conscious states. Every test that we could possibly do, from asking the zombie the price of milk to requesting that he write a poem about birdsong, produces a reaction that is indistinguishable from that of a normally functioning human being of average intelligence, yet mentally speaking, there is 'nothing there'. The big philosophical question here is: are zombies really conceivable? The description just given, however implausible, seems intelligible: it does not appear to contain any contradiction. But if so, this leads to a disturbing thought. What if other people are really just zombies and I am the only conscious being around? Of course no normal person can believe this in everyday life. But am I philosophically justified in rejecting it?

Actually, we have already indicated a possible solution to this problem in the form of the argument from analogy (see above). We have also explained why this solution tends these days to be rejected, largely as a result of Wittgenstein's influence. (Wittgenstein would have said that zombies were not conceivable in the first place, as he rejected the sharp distinction between what is publicly observable in creatures' behaviour and what is going on 'inside their heads'.) But what I want to draw your attention to at this point is the connection between the zombie idea and the issue of qualia that we were discussing earlier. If you are a physical reductionist and think that even qualia can be reduced to brain states you can dismiss zombies as impossible in principle. This is because you will think that once it is established that the beings alleged to be zombies are physically and behaviourally just like human beings, then there is, so to speak, nothing missing from their humanity: they must be human persons. In relation to people, the physical reductionist accepts the adage 'If it looks like a duck, swims like a duck and quacks like a duck, then it must be a duck' (but bear in mind that this a slogan and not an argument!) and because of this she need not worry (or at least thinks she need not worry) about the 'problem of other minds'. Of course, the question still remains of whether we are entitled to take this view, given the difficulty of rejecting the separate reality of qualia.

**Non-reductive materialism**

**1. Supervenience and emergence**

If we still want to be materialists but are persuaded that reductionist versions of materialism cannot be made to work satisfactorily, what are the alternatives?
In order to understand the various options, it useful to introduce two new concepts: *supervenience* and *emergence*.

Supervenience means that if you duplicate all the lower-order facts, then you duplicate the corresponding higher-order facts as well. So, for example, if consciousness supervenes on the brain, then any two brains in precisely the same physical state over a given period would be having the same conscious states during that period. In all forms of materialism (apart from eliminative materialism, which does not recognise the existence of consciousness at all) consciousness supervenes on the brain.

*Emergent* properties and processes are those that arise from the interactions between the lower-order entities and cannot be deduced from the properties of those entities considered in isolation. If consciousness is correctly described as an emergent property of the brain, this implies that it arises from neurons, but that it cannot be deduced from the properties of neurons considered in isolation—their interactions also have to be considered. (An example of a non-emergent property of a brain might be its mean density, as this is purely determined by the masses of the different bits of matter in the brain and their distribution—their interactions are irrelevant).

In addition, we can distinguish between weak and strong emergence.

A property is only weakly emergent when, in principle at least, it can be deduced from all the lower-level facts (including facts about interactions). It is strongly emergent when it cannot be so deduced. According to Australian philosopher David Chalmers (b. 1966), consciousness is a strongly emergent quality relative to brain facts, because you can't deduce it merely from the physical properties of a brain (even properties related to interactions) that it is conscious. In order to make such a deduction you would have to add laws that correlate conscious states with brain states, laws that are quite unlike those of standard science, as they involve a non-physical component.

The distinction between weak and strong emergence can also be used to explain further the difference between reductive and non-reductive versions of materialism. In reductive materialism, the mind is only weakly emergent, whereas in non-reductive versions (with the exception of eliminative materialism, which denies the existence of the mind altogether) the mind is strongly emergent, as it cannot be deduced from the physical facts about the brain.

2. Anomalous monism

Anomalous monism, which was developed by American philosopher Donald Davidson (1917-2003), is one of the hardest theories of the mind to grasp. This is because it denies the existence of precise laws linking the mental and the physical *and nevertheless* claims that mental states are actually physical. Two things therefore need explaining. First, why does Davidson think that there are no precise psychophysical laws? And secondly, why, despite this, does he think that mental states are physical?
Davidson thinks that precise laws linking mental states with physical states are not possible for the following reason. Suppose we were trying to figure out the mental states of Martians, i.e., what they think and feel in different situations. Obviously, we would have to infer this from their observable behaviour. But, Davidson argues, the behaviour would not provide enough information. We need to be able to interpret that behaviour and for this, we need to assume—without any prior evidence—that it is rational. For example, if a Martian puts on his breathing apparatus when arriving on Earth, we should assume (absent any evidence to the contrary) that the Martian does this because he believes that he cannot breathe Earth's atmosphere and he does not want to die. (The words 'believes' and 'want' are italicised here to draw attention to their mental status.) Thus, Davidson argues, our attribution of mental states to an alien being of this sort is subject to the fundamental constraint that it should be compatible with assuming that the alien is rational. Now, Davidson continues, when scientists attribute physical states to another being (e.g. a particular type of brain state to a human being), they do not apply this, or any similar, constraint. But if there is a certain constraint that governs our mental ascriptions and no parallel constraint governing our physical ascriptions, then there cannot be precise laws linking the occurrence of particular types of physical states with the occurrence of particular types of mental states. This is the first thesis that Davidson wanted to establish.

The second thesis was that, although (as just shown), the mental and the physical cannot be correlated or identified at the level of types, they can, and indeed must, be identified at the level of tokens. That is to say, we should recognise that individual mental states are identical to individual physical states. This is because our mental states such as beliefs and desires can cause physical events and be caused by them and to say this is, according to Davidson, equivalent to saying that individual mental states are physical. In this way, Davidson is able to reconcile physicalism (materialism) with the impossibility of precise psychophysical laws.

Note that in anomalous monism, the mental is regarded as supervenient on brain states, but also strongly emergent with respect to them, since the impossibility of psychophysical laws would seem to make it impossible in principle to deduce mental states from physical ones.

3. Property dualism

A more widespread theory is known as property dualism. We have already discussed Descartes' theory of substance dualism that the mental and physical constitute distinct things or substances. The difference between the two theories lies in the nature of the 'mental' side. For a substance dualist, it is an immaterial mind or person, which is the possessor of mental properties (such as feeling pain), with a clear distinction between the properties and the mind that possesses them. For the property dualist, however, the world consists entirely of physical (material) objects, but some of these objects have, in addition to their physical properties, mental properties which cannot be reduced to them. A key advantage of property dualism is that, as it only recognises one type of 'stuff' (with both mental and physical properties), it may be considered a simpler and more unified theory than substance dualism.

On the other hand, property dualists, like substance dualists, may have problems with mental causation. Suppose I decide to switch on a light and because of this (as we would say) my arm
moves towards the light switch and flicks it down. As we saw, the substance dualist has difficulty in explaining how the mental act of my deciding to switch on the light interacts with the physical neural cells that initiate the relevant chain of events. Does the property dualist also face this problem or does she avoid it? Well, for the property dualist there is one event, which has both mental properties (e.g. being a decision) and physical properties (e.g. being a set of neural firings). There is no problem about this event leading to a physical process that brings it about that the light is switched on. It can do this because it has physical properties. But is this enough to secure our intuition that through our minds, we are able to affect our bodies and the world around us? Arguably not, since we would presumably need it to be the case that the event has these effects in virtue of its mental properties. But there appears to be no reason to say this, as a causally complete story of what happens can be given that makes reference only to the physical properties. In other words, we seem to be back with a kind of epiphenomenalism, in which the mind is just a by-product of physical processes and cannot actually affect them in any way.

4. Biological naturalism

Biological naturalism is a non-reductive theory of consciousness developed by the American philosopher John Searle (b. 1932). According to Searle, consciousness is as much an ordinary biological phenomenon as digestion. But like the property dualists, he regards consciousness as strongly emergent and so not reducible to matter. He maintains that consciousness is caused by brain processes, but not that it is an effect that is somehow distinct from those processes, as this would imply dualism, which he rejects. (However, by dualism he presumably means here ‘substance dualism’, the idea that the mental and the physical are two distinct sorts of things—he should approve of property dualism, which seems hardly different from his own theory.) He thinks that consciousness is caused by the brain in something like the way the liquidity of water is caused by the behaviour of H₂O molecules. (The former is not a ‘thing’ distinct from the latter.) Biological naturalism therefore has the advantage of ontological simplicity, since it maintains that everything is physical. But again, we may feel doubts regarding the extent to which the theory succeeds in explaining consciousness. How convincing is the analogy with water molecules and the property of liquidity? One problem is that, given the structure of water molecules and the nature of the links between them, their possession of liquidity seems to be in some sense necessary, whereas the relationship between the brain and consciousness appears to be a contingent one, as the zombie problem shows. (There could conceivably be living, fully functioning human bodies and brains that were not conscious—though, as we saw, not everyone agrees with this.)

5. Eliminative materialism

This is perhaps the most radical theory of all. It says (to put it baldly) that the mind simply doesn’t exist. It seeks to ‘eliminate’ the mental from our conception of the world, which, it maintains, should be thought of as containing only material or physical things. Some people believe in demons, but it does not follow that there are any such things in the world. Likewise, says the eliminative materialist, with minds.
At first sight, such a view may seem incredible. Is it not just a simple fact that people believe things, understand things, experience sensations, dreams and so on? How could we be wrong in supposing such things to be true? Well, according to the eliminative materialist, such talk represents a kind of theory, called folk psychology. We do not immediately perceive beliefs, desires and other standardly recognised mental phenomena. These are things that we ‘posit’ (i.e. suppose to exist) in order to predict and explain what we do observe (behaviour). In principle, a theory can always be mistaken—the things posited by the theory might not exist (just as the substance known as ‘phlogiston’, posited by early chemists to explain burning, turned out not to exist). So it might turn out that beliefs, desires and the rest do not exist, despite the fact that they are so prominent in our thinking about our own and others’ behaviour. But are there any reasons for thinking that folk psychology is mistaken?

According to Canadian/American philosophers Patricia and Paul Churchland (b. 1943 and 1942 respectively), there are plenty. Here are some of the problems alleged by the Churchlands and other eliminative materialists:

- Folk psychology leaves many aspects of our mental lives unexplained: it cannot explain such things as dreams as well as various aspects of memory, learning, consciousness and mental illness.
- Beliefs, desires and other so-called propositional attitudes have a certain structure which has no counterpart in the brain. For example, believing is always believing that something or other is the case, where the ‘something or other’ is expressed by a sentence. The sentence is made up of symbols which are combined together according to certain grammatical rules (or syntactic rules, as they are also called). But brain processes involve no such things. When describing events in the brain, physiologists talk rather of the electrical and chemical properties of brain cells.
- There also seems to be no counterpart in the brain of the semantic properties of propositional attitudes, i.e., those that relate to meaning. It is pointed out that in folk psychology we normally individuate psychological states in terms of what they are about. But this is often vague and breaks down altogether in the case of the mentally ill and the very young, whose understanding of meanings is limited.

Here are some responses to aspects of the above arguments:

- It may be true that folk psychology does not explain everything, but it hardly follows that it must be totally rejected. After all, we do commonly use it successfully on an everyday basis to make sense of our own and others’ actions.
- Those who argue that there are no neural counterparts of such things as symbols, syntactic structures and meanings are invoking the wrong level of analysis. It is equally true that when we look inside a computer we do not see any software, data or programs. (We see only boards of microchips, wires and so on.) But that is only because the programs and data do not exist at that level. Similarly, those who are looking for symbols and meanings inside the human brain may be confused about levels—these things exist at a higher level of analysis than that of nerve cells.
As well as criticising talk of propositional attitudes such as beliefs and desires, some writers have taken exception to qualia (see above). Daniel Dennett has argued that there is nothing that has the properties typically attributed to qualia. Traditionally, those who have believed in qualia have supposed that we have infallible knowledge about own current qualia. But in the case of pain qualia, they have also supposed that they must be intrinsically awful. Dennett argues that they cannot have both properties because of the phenomenon of reactive dissociation: under certain conditions, patients who have been given drugs like morphine report feeling genuine pain but not minding about it. So pain is either not intrinsically awful or it is possible to be mistaken about whether we are in pain.

Some of the problems of eliminative materialism are common to reductionism. In particular, if the knowledge argument (see above) is sound, then it works just as well against eliminative materialism as it does against reductionism. (If Mary in the story, despite knowing all the physical facts, doesn’t know what it is like to see green, then there is some real aspect of the world—the mental aspect—which she doesn’t know about). In fact sometimes it is not easy to distinguish the two theories. Basically, an eliminativist will say that the mind does not exist, whereas a reductionist will say that it exists but it is actually just physical. It might be argued that reductionism gives a slightly less radical gloss to what is essentially just the same theory as the eliminativists hold.

**Artificial intelligence**

*Artificial intelligence* refers to the ability (real or imagined) of some machines to think, in the same sense of the word that human beings can think. The philosophical controversy surrounding artificial intelligence (AI) is about whether it is possible or whether thinking is the preserve of human beings and perhaps of a few other animals.

Philosophical treatment of the issue of AI begins historically with the publication of a famous paper in the journal *Mind* in 1950. The paper was entitled 'Computing Machinery and Intelligence' and it was written by the British mathematician and wartime cryptographer Alan Turing (1912-1954), who had himself already described the basic structure of the modern digital computer. In this paper Turing makes a proposal as to how we might come to know whether a machine could think. His suggestion has become known as the *Turing Test* and in essence it comes down to this: if you are conversing with something and ask it anything you like and you cannot distinguish at all between its responses and those of a normal human being, then you may conclude that this thing that you are conversing with is able to think.

Turing's proposal is controversial. This is mainly because it is an entirely *behavioural* criterion of whether an entity has the capacity for thought: like physical reductionism, it applies to the attribute of thinking the adage: 'If it looks like a duck, swims like a duck and quacks like a duck, then it must be a duck'. But many people believe that there must be more to thinking than just the ability to 'say the right kinds of things'. Thinking, they would argue, must involve consciousness and, as the knowledge argument appears to show, there is more to consciousness than just the physical facts. On this view, however good a machine's *performance* is, it remains an open question whether it is actually thinking, because it is an open question whether it is 'like anything
to be' the machine. Most of us have been tempted to say things like 'My computer thinks it has already saved the file' and, according to the view we are considering, such talk might never literally be true--however sophisticated the computer was.

There are consequently two key questions that need to be carefully distinguished in this area. One is: "Is it the case that for any possible ‘thinking task’ that can be carried out by a human being, it is possible in principle to make a machine that can do this task so well that it could pass the Turing test?" The other is: "Could a machine in principle be designed to have a fully-fledged ability to think, such that it could be said to understand what it was doing?"

For both questions, the answer is far from obvious. Many people doubt that a computer could ever pass the Turing test, on the grounds that there are some tasks--such as the writing of love poetry--which are simply beyond the capacity of a machine. But as time goes on, it becomes harder and harder to sustain this view in the face of the evidence. Computer programs have been written which write pretty interesting poetry. In fact, the range of tasks that computers cannot do gets steadily smaller as time passes. But I think there is an a priori reason why we should think any human task reproducible in principle by a machine. This can be seen by asking how, in very general terms, human beings themselves do any task. Arguably, the correct answer (because it appears to be the only imaginable answer) is that they do it computationally, that is, by means of (often very complex) mechanical processes, and every such process is, by definition, within the power of a machine. Now it might be argued that purely mechanical processes could not be sufficient for every conceivable task, since to suppose this would be to overlook the importance of creativity. This feature seems to be essential to many human tasks from writing poetry to planning a military campaign. Surely, it will be argued, a machine could not display genuine creativity.

But what exactly is creativity? Daniel Dennett has adopted an interesting approach to this issue. His analysis is based on the fact that there are two distinct elements within creativity. One is simply the production of new ideas or strategies at random; the other is the filtering of these ideas through a process that can distinguish more promising ideas from less promising ones. The latter, Dennett claims, is purely mechanical--it proceeds purely by means of rules that could always in principle be applied by a computer. The former is just a matter of chance and it is a straightforward matter to implement--or at least simulate--this in a computer. Hence, Dennett reasons, there is nothing in creativity per se which cannot be realised in purely physical terms. To me, this would seem to support the idea that genuine creativity is within the power of a sufficiently sophisticated machine.

If we accept this, then the answer to the first of our two questions--whether a computer could in principle pass the Turing Test--would seem to be ‘Yes’. But there is still the second--and arguably more difficult--question: could a machine literally understand anything?

To this question, John Searle has given a resounding 'No'. In support of his position, he has devised a thought experiment that has become known as 'The Chinese Room Argument'. The idea (slightly simplified) is this. Imagine you have been locked alone in a room with three documents. The first two documents are written in Chinese, a language which (we assume) you cannot speak or understand. (We would call the first document a 'story' and the second 'questions
about the story', but you don't know this.) The third document is a set of instructions in English telling you how to relate symbols in the second document to symbols in the first one purely on
the basis of their shapes. (We would say that the rules are purely syntactic, not semantic: they are
about the manipulation of symbols, not meanings). In effect (though again you don't know this)
this book of rules is telling you how to 'answer the questions, based on the story'. Using this third
document, you produce a fourth document called 'answers to the questions', which you submit on
leaving the room.

Now let’s suppose you have done this. And let’s also suppose that this document can be
understood as correct answers (in Chinese of course) to the questions that the second document
can be interpreted as asking. Does it follow that you can understand Chinese? Of course not. It
was stipulated that you knew nothing of this language—or at least nothing of the meanings of the
words and sentences within it. You were able to do the job only because you were good at
following instructions about how to manipulate symbols according to syntactic rules. The
corresponding point about AI programs and the computers running them is this: these programs
use nothing but ‘meaningless’ rules of precisely this kind. So shouldn’t we say exactly the same
about their host computers, however sophisticated they may be, namely, that they have no
understanding of anything—and hence do not literally think, as thought necessarily requires
understanding? The conclusion that Searle’s thought experiment is meant to lead us to is that we
should say exactly this.

But is the argument convincing? It depends crucially on how the elements of the two scenarios—
the Chinese room and the AI program and computer—are matched up. Note that in the former
case, you, the occupant of the room, are just one component of the situation. There are also the
documents, including the instruction manual, the pens, the desk and whatever other
paraphernalia you need. What is the analogue of you in the AI system? It cannot be the whole
system: it must be just one part of it. Perhaps it is the computer’s processor, which applies a
program (stored in its memory—this corresponds to the instruction manual) to the data (also
stored in memory—corresponding to the story and questions). But would anyone say that a
processor, when considered in isolation from its stored program and data, understands anything?
Without software, hardware has no ability to work anything out and therefore presumably has no
‘understanding’. If this is right, then the AI computer system as a whole should be compared, not
to any individual element in the Chinese room (the human occupant or any other) but to the room
as a whole, together with its inputs (the first three documents) and output (the fourth document).
The question should be then: does the Chinese room understand anything?

It might be thought that this is hardly much help to the defender of the idea of genuinely
intelligent machines, since it is less plausible to attribute understanding to a room than to a
person! But this would be to miss the point. The Chinese room works like a machine—in fact, it
is a machine for answering questions in Chinese about a story written in Chinese. So if you were
to decide, based on intuition alone, that it cannot understand anything just because it is a room
(with a relevantly ignorant person inside), then you would presumably also have to declare on
the same basis that a computer could not understand anything—it is intuitively no more absurd to
attribute understanding to a room that functions as a machine than it is to attribute it to a bunch
of circuits, wires, plugs, software and so on. But then if you reject the idea of an intelligent
machine on these sorts of intuitive grounds, then you are begging the question against intelligent machines--and wouldn’t even need to use the Chinese room argument in any case.

I would argue, then, that the Chinese room argument fails to show that there cannot be intelligent machines or machines that genuinely think and understand what they are doing. However, in a way, we are left here with more questions than answers. For if we agree that a machine’s performance, however sophisticated and effective in its task it may be, cannot alone settle the question of whether it thinks, then we are still left asking what in general does definitively settle this. Are we forced to bring in what Ryle would describe as the ghost in the machine and which some might call the soul? The issue seems as obscure as ever.
Bibliography


René Descartes, Meditations on First Philosophy, with Selections from the Objections and Replies (trans. Michael Moriarty), OUP (2008).


