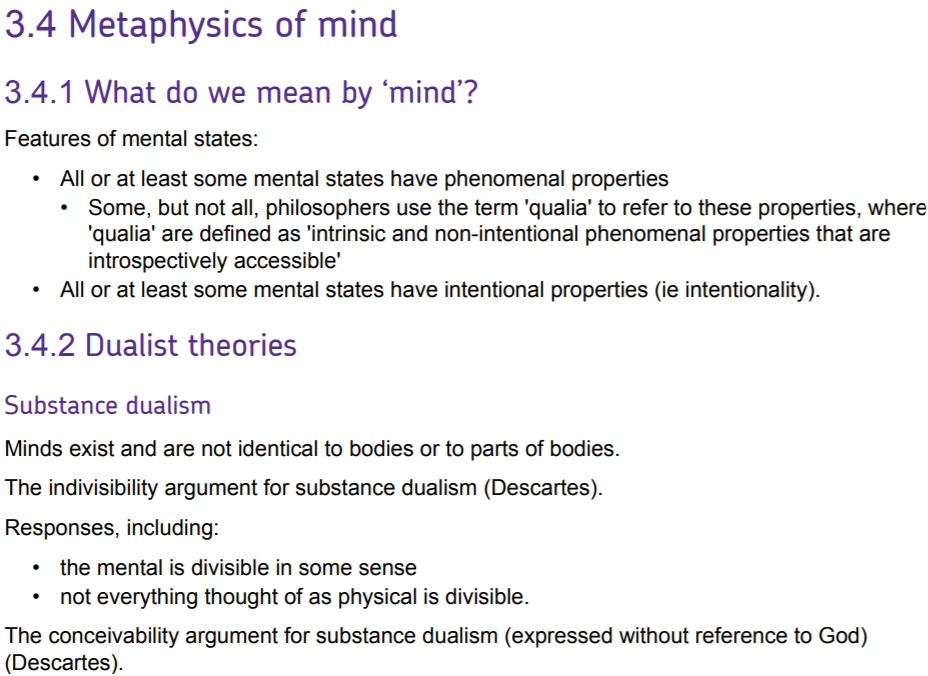
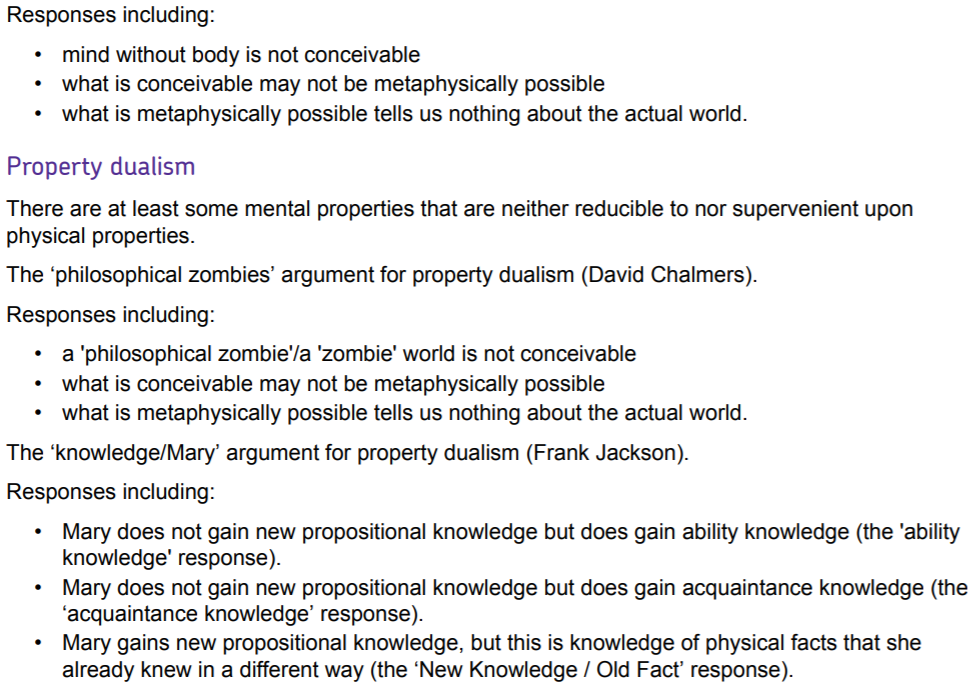
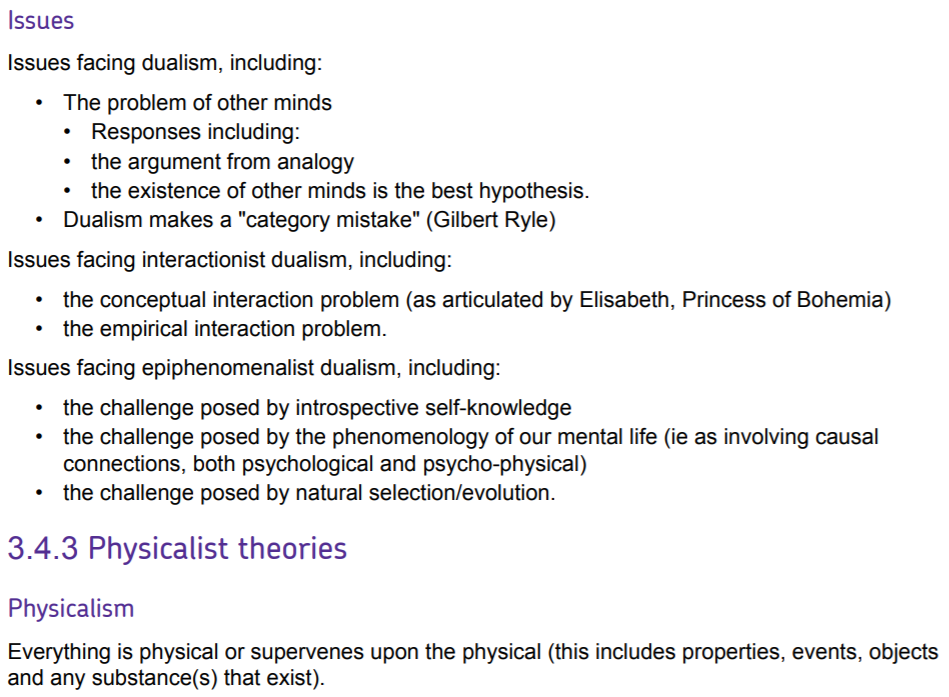
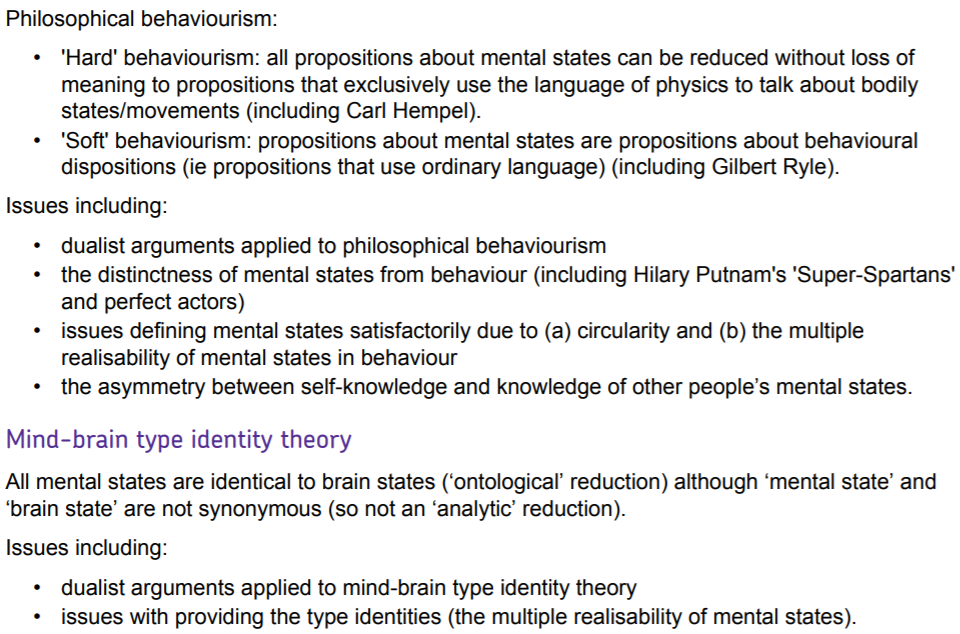


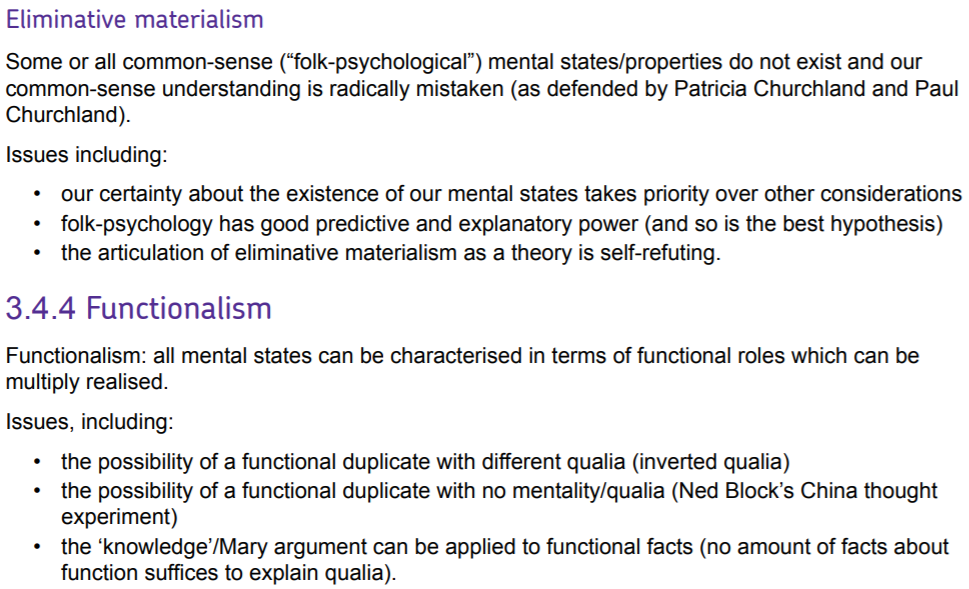
Metaphysics of mind











**1. Metaphysics of Mind: Overview**

Philosophy of mind is a branch of metaphysics, and different theories in philosophy of mind disagree on metaphysical questions about what exists and its nature. Questions about what exists are questions about ontology. It is worth having a sense of the different theories and how they relate to each other at the most general level.

According to a traditional metaphysics, a substance is an entity, a thing, that does not depend on another entity for its continued existence. It has ‘ontological independence’. For example, this book is a (physical) substance. Substances are also understood by contrast with properties.

Substances are what possess properties. The chair (substance) is solid (property). Properties can’t exist without substances – they depend on substances to exist. Solidity depends on things being solid; the property ‘being 1 metre long’ depends on something being that long; and, Descartes claimed, thoughts can’t exist without a thinker.

Substances persist through changes in properties – something can change from being 1 metre long to being 1.1 metres long, e.g. by growing. Obviously, its property of ‘being 1 metre long’ does not persist through this change. It loses that property and gains another. Or again, a thinker can think a series of thoughts – the thinker persists, the thoughts do not.

A central question in metaphysics of mind is ‘is the mind a substance?’ Can your mind exist on its own, independently, or is it dependent on something else in order to exist? In particular, is your mind dependent on your body, perhaps especially your brain, in order to exist at all? Many people believe, and many religions teach, that your mind can exist after death, i.e. the death of your body. This can mean many things, which we can’t review here, but one common interpretation is that your mind is a separate substance from your body. If the mind is a substance, then the end of your body’s existence is not the end of your mind’s existence.

The view that the mind and the body are separate substances is known as substance dualism. Substance dualism claims that there are two fundamental kinds of substance – mental and physical.

We can contrast substance dualism with idealism, the view that minds are the only kind of substance, and so whatever exists is either a mind or depends on a mind. However, the most popular alternative to dualism is the view that the only kind of substance is physical. On this view, because the mind is not a separate thing, it is more accurate to talk about mental properties. These are often understood to be properties of Intentionality and consciousness (for more on these features of mind, see the section ‘What do we mean by ‘mind’?’). Mental properties include mental states, such as beliefs, and mental events, such as having a thought or feeling pain. If the only kind of substance is physical, then mental properties are properties of a physical substance. You can think of the physical substance that has such properties as either the person or the brain. If substance dualism is right, then mental properties are possessed by mental substances.

Suppose there is only physical substance. Our question about the mind now becomes a question about mental properties and their relation to physical properties. Are thoughts and conscious experiences themselves types of physical property (physical states and events)? There are many different types of physical property, e.g. size, shape, motion, mass, various forms of energy, chemical properties such as molecular structure, biological properties such as genetic code, and many others. Are mental properties, such as thinking about snow or feeling sad, also physical properties? One might argue, for instance, that they are simply neurological properties, e.g. to think about snow just is for certain neurons to fire in one’s brain. This view is known as the mind–brain type identity theory.

Alternatively, one might argue that mental properties are not physical properties and can’t even be understood or explained in terms of physical properties. (Neurons firing is just a physiological process, like food being digested. How can consciousness be neurons firing any more than it could be digestion?) Although mental properties are possessed by physical substances, they are completely different from any of the other properties physical substances possess. This view is property dualism – there is only one kind of substance but two radically different kinds of property.

On the other hand, if we can’t say that mental properties are physical properties, then perhaps we should question whether there are any mental properties as we usually think of them. Eliminative materialism claims just this. Our whole common-sense way of thinking about ‘the mind’ and mental properties is mistaken, and we shall need developments in neuroscience to provide a new way of explaining human beings.

Another approach argues that all the views above are misled in their metaphysics. Mental ‘properties’ aren’t like physical properties, ‘only different’. Instead, talk of the mind, of thought and consciousness, should be understood in terms of behaviour and dispositions to behave. Wanting to go for a drive, feeling cross, thinking about your mother – these are each a matter of being disposed to behave in certain, perhaps highly complex, ways. This view is philosophical behaviourism.

Our final theory follows this lead, but argues that thinking of mental properties just in terms of behaviour is too restrictive. We can talk more broadly in terms of the contribution of mental properties to how the person (or brain) functions, including the interactions of mental properties with each other, such as how one thought leads to another, how desires lead to emotions and vice versa, as well as how any and all of these lead to behaviour. This view is functionalism.

**1.1. What do we mean by ‘mind’?**

What are you? I mean, what is it to be you? What kind of thing is a human being? What makes a person a person? People have given surprisingly different answers to these questions. You might think, in light of evolutionary theory, that the answer is that we are animals. But sometimes, when someone is an irrecoverable coma or brain-dead, we say that they no longer exist, that they’ve ‘gone’. But the body lying there is still the same animal. So our minds seem particularly important to who or what we are. Without a mind, I am not a person at all and I’m not ‘me’.

But what do we mean by ‘mind’? Many philosophers think that the mind has two important features, ‘thought’ and ‘consciousness’. These terms pick out the two most important aspects of what we mean by ‘having a mind’. We can talk about ‘minds’ or about ‘mental properties’, to include mental states, like beliefs, and mental events, like thinking a thought or feeling a pain. What distinguishes things that have minds or mental properties from things that don’t is that things with minds have a ‘point of view’, a ‘perspective’, on the world. Things with a point of view experience the world, there is a ‘subjectivity’ to their existence, they are not just objects. And this involves ideas of being conscious and of being able to experience and think about things, to have beliefs and desires.

**Intentionality**

Thoughts are ‘about’ something, objects or events in the world. For example, I might have a belief *about Paris*, a desire *for chocolate*, be angry *at the government*, or intend *to go to the pub*. In all these cases, my state of mind is ‘directed’ towards an ‘object’, the thing I’m thinking about (Paris, chocolate, the government, going to the pub). This idea of ‘directedness’ is known as ‘Intentionality’.

Intentionality is not about intentions (to mark the difference, I shall use a capital ‘I’ for ‘Intentionality’). If I have an intention, I am ‘aiming at’ *doing* something. With Intentionality, it is the thought or mental state which ‘aims at’ its object, what it is about, and no ‘doing’ needs to be involved. Beliefs, desires, emotions all have Intentionality; they are all about or concern some object or other. They are all ‘Intentional mental states’.

Whenever we think of, have a belief about, or desire something, we always conceive of it in a certain way, under a particular description. For example, in Sophocles’ famous play *Oedipus Rex*, Oedipus kills his father and marries his mother. He doesn’t want to do this. But he doesn’t know who his (biological) parents are. On his journeys, he meets an old man in the road who gets in his way. Oedipus becomes very angry, and kills the old man. In fact, the man was his father, Laius. Oedipus was angry at the old man. Was he angry at his father? From his point of view, he wasn’t – he didn’t think of the old man as his father. So Intentional states represent the world in particular and partial ways. It’s like seeing something – a desk, say – from a particular aspect; you can see it, but not all of it.

What Intentional states represent – Paris, chocolate, the government, going to the pub, Laius – is called the ‘Intentional object’. The way they represent that object we can call the ‘aspectual shape’ of the object. The Intentional object + the aspectual shape comprise the Intentional content of a mental state. The Intentional content of a mental state is the answer to ‘what are you thinking (about)?’ The way the person answers the question will also tell us how they are thinking about it, e.g. ‘I’m angry at the old man’.

We can now say that an Intentional mental state is a mental state with Intentional content. But we can add to this. We can have different mental states with the same Intentional content if we take different ‘attitudes’ to that content. For example, I can believe I’m arriving late; I can want to be arriving late; I can fear I’m arriving late; I can be pleased I’m arriving late. An Intentional state, then, comprises a particular ‘attitude’ or ‘mode’ towards a particular Intentional content. (Many philosophers calls these mental states ‘propositional attitudes’, because the Intentional content is (usually) expressed as a proposition.)

It is a debate in the philosophy of mind whether *all* mental states have Intentionality and whether there are *other* mental properties besides the properties of Intentionality. One aspect of that debate concerns how we should best understand the second feature of the mind, consciousness.

**Phenomenal properties/qualia**

Consciousness, especially the sort of consciousness involved in perception, sensation and emotion, has a ‘feel’ to it, a distinctive ‘experiential quality’. The phrase often used to try to capture this experiential quality is ‘what it is like’. There is something it is like to taste beer, to see a red rose, to feel sad.

‘What it is like’ here isn’t meant to compare the experience to other experiences, it is meant to pick out how the experience is for the subject. When we make comparisons between experiences (e.g. ‘Seeing a red rose is like seeing a ripe tomato’), we do so *in virtue of* what it is like to see a red rose in the sense meant here. It is the experience of redness that allows us to compare roses and tomatoes; and there is something it is like to experience redness. Similarly, there is something it is like to feel sad. I don’t mean by this that feeling sad is like feeling some other emotion. I mean that there is a distinctive ‘feeling’ to sadness.

We can call the properties of an experience which give it its distinctive experiential quality ‘phenomenal properties’. We are aware of these properties through consciousness and introspection, by turning our attention to our conscious experiences themselves.

Some philosophers call phenomenal properties ‘qualia’. However, to do this usually means that the philosopher has a particular *theory* of phenomenal properties in mind. Phenomenal properties are only qualia if they are *intrinsic*, *non-Intentional* properties of experience. What does this mean?

An intrinsic property is one that its possessor (in this case, the experience) has in and of its own, not in virtue of its relations to anything else. Think of the smell of coffee. It is the smell ‘of coffee’ because of its relation to the substance of coffee. That it is ‘of coffee’ is not an intrinsic property. But consider: something else could cause the *same* smell as the smell caused by coffee. So, someone who believes in qualia would argue, what makes that smell the smell that it is, is not the fact that it is caused by coffee. How that smell smells is an intrinsic property, because it would be that smell even if it wasn’t caused by coffee. The smell can’t be analysed just in terms of what causes it. Another example: pain wouldn’t be pain if it didn’t *feel* painful, whatever it is or isn’t caused by. Phenomenal properties of experience, then, are intrinsic, and their identity is fixed not by what causes the experience (or what the experience causes) but by how the experience is in itself. Or so people who believe in qualia argue.

Intentional properties, we saw above, are properties of a mental state that enable it to be ‘about’ something, to represent what it does. So Intentional properties are a matter of how the mental state ‘hooks up’ to the world. So they are relational rather than intrinsic properties. That a belief is about Paris is a property it has in virtue of its relation to Paris; that a desire is for chocolate is a property it has in virtue to chocolate; and so on. So qualia, because they are intrinsic properties, are non-Intentional properties.

Some philosophers also argue that qualia are, by definition, ‘introspectively accessible’. Introspection is just turning your attention to your own mind. Because qualia are properties of consciousness, they are properties we are consciously aware of. What it is like to smell coffee is something that we can know by being conscious of the smell of coffee. So we can ‘access’ qualia through introspection, by consciously turning our attention to our own conscious experiences.

But do qualia, in this more specific sense, exist? Or put another way, are phenomenal properties of consciousness qualia? Are they intrinsic and non-Intentional, or can they be analysed in terms of their Intentionality? For example, isn’t the smell caused by coffee the smell *of* coffee? Doesn’t this smell represent, isn’t it ‘about’, coffee? People who believe in qualia would argue that it isn’t. The smell that is caused by coffee isn’t – in and of itself – *about* coffee, because exactly the same smell could be caused by something else. When we say that it is the smell *of* coffee, we add to the experience itself a thought about what caused the smell – and it is this thought, not the smell itself, that has Intentional content that links the smell to coffee. Our conscious experiences have an element of Intentional content, they also have an element – qualia – that is non-Intentional and intrinsic. We cannot, for instance, understand and explain how red looks to us or what it is like to feel pain just in terms of what these mental states are about, how they are related to the world.

This claim about the relationship between phenomenal properties and Intentional content is controversial. And that is why the debate over whether phenomenal properties are qualia – i.e. whether phenomenal properties are intrinsic, non-Intentional properties, or whether they are properties of Intentional content – continues.

**2. Dualism**

**2.1. Descartes’ indivisibility argument for substance dualism**

Substance dualism claims that both minds and bodies – physical objects – exist. It is common in contemporary philosophy of mind to assume that bodies exist, and we shall share that assumption. Substance dualism is controversial, therefore, in claiming that the mind is an ontologically distinct substance.

Substance dualism holds that there are two fundamentally different types of substances. In traditional dualism, these two types of substances are physical substances (‘bodies’, physical objects) and mental substances (minds). Minds are distinct from bodies – they are not bodies, they are not parts of bodies, and because they are substances, they are not properties of bodies either. Cartesian dualism – the form of substance dualism defended by Descartes – also claims that minds do not depend on bodies in order to exist, i.e. minds can exist separated from any body. People who believe that the mind is the soul, and the soul can continue to exist without a body after death, are usually substance dualists.

If mental substances exist, they will be very unlike physical substances. In this section, we discuss Descartes’ argument that minds are distinct from bodies because they do not have any parts and do not even exist in space.

**Descartes’ indivisibility argument**

In Meditation II, Descartes argues that mind and body have different essential properties – thought and extension. He understands thought in terms of consciousness and Intentionality; extension is the property having a size and taking up space. In Meditation VI, Descartes claims that this provides an argument that mind and body cannot be the same thing: if they were the same thing, they would have the same properties.

Leibniz later formalised this claim in his principle of the indiscernibility of identicals: if two things are identical (i.e. are just one thing), then they share all their properties. Why? Because one thing cannot have different properties from itself. So if two things have different properties, that proves that they cannot be one and the same thing.

But why think that the mind has different properties from the body? Descartes argues that, unlike physical objects, the mind does not have any parts and cannot be divided, and so it is not extended:

*“When I consider the mind – i.e. consider myself purely as a thinking thing – I can’t detect any parts within myself; I understand myself to be something single and complete … the faculties of willing, of understanding, of sensory perception and so on, these are not parts of the mind, since it is one and the same mind that wills, understands and perceives.”*

Willing, understanding and perceiving are properties of the mind, different ways of thinking. By contrast, the body does have parts. You can literally lose part of your body, e.g. a hand. So the body (physical substance) is divisible into parts, but the mind (mental substance) is not. So mind and body are distinct types of thing.

**The mental is divisible in some sense**

Descartes’ claim that we will, think, imagine, with the whole of our minds, not a literal part, is appealing. However, cases of mental illness, e.g. multiple personality syndrome, might be used to suggest that the mind can be divided. In such cases, it seems that some ‘parts’ of the person’s mind are unable to communicate with other ‘parts’. Theories of the unconscious suggest something similar: people may believe or desire one thing consciously and the opposite thing unconsciously. So it makes sense to talk about ‘parts’ of the mind.

However, Descartes could respond that the way in which the mind is ‘divisible’ is entirely different from the way in which the body is. Bodies are spatially divisible, while minds are only functionally divisible. The different ‘parts’ do different things, but they aren’t in different spatial locations. So his argument that mind and body are different because they have different properties is unchallenged.

**Not everything thought of as physical is divisible**

Is it true that physical substances are always divisible? When Descartes argued that extension is the essential property of physical objects, and that what is extended is divisible, was he right? We may question whether this theory of physical objects is correct. It was a matter of some debate in the seventeenth and eighteenth centuries whether physical objects are infinitely divisible. If you cut something up, can you always cut it into smaller pieces? The question is not whether we can actually do this right now, with the technology we have, but whether there are physical things that cannot be divided even in principle. If, for example, the smallest physical particles are best understood as packets of energy or force fields, then we can’t further divide these – you can’t have half a force field! Or again, perhaps not only force fields but also processes or waves or something else that can’t be divided spatially form a fundamental part of the physical universe.

One possible response is that even if force fields or waves can’t be divided in reality, we can still conceive of them having half the size. In that sense, we can still talk of spatial ‘parts’. There is no logical limit to how small spatial parts can be. However, whether this is true or not may depend on the best physical theory of what space is. If we need to change our concept of space, then perhaps there will be such a limit.

The implication of these reflections is that it may not be an essential or defining property of every physical substance that it is divisible. There are some indivisible things, such as force fields, that are physical. In that case, the fact that the mind is not divisible does not entail that it is not physical. It could be a form of non-divisible physical thing. So even if Descartes is right that the mind isn’t divisible, this doesn’t prove that it isn’t physical.

This line of thought does not show how the mind could be a non-divisible physical thing. After all, the mind is very different from subatomic particles! The objection only seeks to show that Descartes’ divisibility argument, as it is stated, fails.

**Is the mind a substance?**

Perhaps a more conclusive objection to Descartes’ indivisibility argument is that the argument assumes that minds exist as substances, and then argues that they are not physical substances. But suppose that minds are not ‘things’, not substances at all. Then minds are not divisible or indivisible – they simply don’t exist. Instead, there are only mental properties – thoughts, desires, pains, etc. Perhaps they are properties of the brain. It is true that these properties are not spatially divisible, but that is because properties in general are not divisible. It is only substances that literally have parts. For example, while the brain has spatial parts, the temperature of the brain does not have parts, and yet it is a physical property of a physical substance. Properties themselves don’t ‘take up space’ in the way that physical substances take up space, and so they can’t be divided into spatial parts. But a substance that is spatially divisible can nevertheless possess properties that are not divisible.

In his argument, Descartes cannot assume that the theory of the mind supposed here is false. He needs to show that it is false. To do that, he first needs to show that the mind is a substance, not simply a way of talking about mental properties, and then he can use the indivisibility argument to show that it is not a physical substance.

**2.2. Descartes’ conceivability argument**

In *Meditation* VI, Descartes presents this argument for substance dualism:

P1. I have a clear and distinct idea of myself as something that thinks and isn’t extended.

P2. I have a clear and distinct idea of body as something that is extended and does not think.

P3. If I have a clear and distinct thought of something, God can create it in a way that corresponds to my thought.

C1. Therefore, God can create mind as something that thinks and isn’t extended and body as something that is extended and does not think.

C2. Therefore, mind and body can exist independently of one another.

C3. Therefore, mind and body are two distinct substances.

In (P1) and (P2), Descartes appeals to the concepts of mind and body that he argued for in *Meditation* II. We can understand (P1) and (P2) to entail the claim that it is conceivable that mind can exist without body. Nothing in our concepts rules this out.

In *Meditation* VI, Descartes adds (P3). Assuming that God is omnipotent, the only reason for thinking that God cannot make something is that the concept of it is contradictory. The concepts of mind and body aren’t self-contradictory. So God can create the mind and the body just as Descartes conceives of them – a thinking thing and an extended thing. We can summarise (P3), (C1) and (C2) in terms that don’t refer to God: it is possible that mind can exist without body.

Finally, a substance is something that does not depend on another thing in order to exist. In other words, a substance can exist independently, on its own. This underpins the inference from (C2) to (C3).

We now have a simpler form of this argument:

P1. It is conceivable that mind can exist without body.

C1. Therefore, it is possible that mind can exist without body.

C2. Therefore, mind and body are distinct substances.

It is important for Descartes’ argument that our clear and distinct ideas of mind and body are complete and exclusive. The mind is nothing but thought; the body is nothing but extension. We know this to be true, he says, because the ideas of mind and body are clear and distinct.

**Mind without body is not conceivable**

Many philosophers believe that Descartes’ conceivability argument doesn’t work. Objections to an argument either challenge the truth of one of the premises or they challenge an inference. On the simplified version of the argument, there is only one initial premise, (P1) ‘It is conceivable that mind can exist without body’. There are then two inferences. First, Descartes infers possibility from conceivability – (C1) because it is conceivable that mind can exist without body, it is possible that mind can exist without body. The second inference is (C2) from the possibility that mind can exist without body to substance dualism. We will look at an objection to each stage of the argument. We start by challenging (P1).

Is Descartes right that we can conceive of mind and body as separate substances, that we can conceive of mind existing without body? Or more precisely, is he right to claim that we can do so clearly and distinctly? Descartes assumes that he can identify what it is to think from introspection. But what is thinking, really? What is its nature? If we knew the answers to these questions, we may find that we cannot clearly and distinctly conceive of thought (mind) without the body. Descartes may think that it is conceivable that mind and body are distinct substances when, in fact, it isn’t conceivable. He may be confused or simply lack relevant information.

For example, the theory philosophical behaviourism argues that the mind – mental states and events – should be analysed in terms of behaviour. To talk of beliefs, thoughts, desires, choices and so on is to talk of how something behaves. Now, without a body, something can’t exhibit behaviour; and without behaviour, there is no mind. If this theory is correct, then once we’ve understood what we mean when we talk about the mind, we will realise that mind without body is inconceivable.

This is a very strong claim. For example, if it is right, then disembodied minds, such as God, are inconceivable. And yet for most of the history of humanity, people have claimed to be able to make sense of the idea of God. So, we may object, defending Descartes, that it is likely that philosophical behaviourism is wrong to think that in talking about mental states, we are talking about behaviour. And so mind without body is conceivable.

However, philosophical behaviourism provides just one argument supporting the claim that mind without body is inconceivable. There may be others. The general point is that we can make mistakes over what we think is conceivable.

Descartes accepts this. We can make mistakes, which is why we must get our ideas clear and distinct first. His claim is that we can’t make mistakes with clear and distinct ideas. So to object to the first premise of the conceivability argument, what we actually need to argue is one of two things. Either we cannot clearly and distinctly conceive of the mind as separate from the body – as the analysis of philosophical behaviourism claims. Or we can challenge Descartes’ theory of clear and distinct ideas guaranteeing truth. Perhaps we can make mistakes concerning even what we conceive clearly and distinctly.

**What is conceivable may not be metaphysically possible**

Let us suppose that we can conceive of our minds and bodies as distinct substances. Just because we can, this doesn’t mean that our minds and bodies really could be distinct substances. Perhaps to exist at all, minds must depend on bodies in some way that we don’t know about. This objection challenges the first inference in Descartes’ (simplified) argument.

To understand the objection, we first need to clarify what is it for something to be ‘metaphysically possible’. And to do that, we first need to understand physical possibility and logical possibility.

**Physical possibility**: Call the world we live in, as it is in fact, the ‘actual world’. This world has particular laws of nature, such as the law of gravity and e = mc2, and physical constants, such as the speed of light. These laws and their application to physical objects define what is physically possible. For instance, it is not (physically) possible for human beings to fly unaided (on Earth), because the upward thrust they can generate using their bodies cannot exceed the force of gravity. What is physically possible is what is possible given the laws of nature as they are in the actual world.

**Logical possibility**: Logical possibility is easiest to understand by relating it to analytic and synthetic propositions. All meaningful synthetic propositions describe what is logically possible. True analytic propositions describe what is logically necessary (what must be the case). False analytic propositions describe what is logically impossible (what cannot be the case). For example, it is logically impossible for there to be a square with three sides. The phrase ‘a square with three sides’ is conceptually incoherent, i.e. the meanings of the terms contradict each other, and so no such thing can exist. Anything that is not logically impossible is logically possible (or logically necessary).

So, we can think of logical possibility as conceptual possibility – what our concepts allow as making sense. We can argue that this is the same as what is conceivable – what we can imagine without self-contradiction.

The laws of nature seem contingent, i.e. it seems possible that they could have been otherwise. Light could have gone faster or slower; the ratio of mass to energy could have been e = mc, and so on. Of course, these things aren’t physically possible. But they are, it seems, logically possible. Nothing in the concept of light entails that it must travel at 299,792 kilometres per second. Or again, it isn’t logically impossible that human beings can fly unaided, just physically impossible.

Everything that is physically possible is logically possible (unless our concepts are terribly muddled!), but not everything that is logically possible is physically possible.

**Metaphysical possibility**: Some philosophers want to stop there, with two types of possibility – physical and logical. But debates in metaphysics, including the metaphysics of mind, over the last 40 years have led many philosophers to argue that there is a third type of possibility, metaphysical possibility. The reason is that analytic truths and necessary truths may come apart.

For example, ‘water’ and ‘H2O’ are different concepts, and before the discovery of hydrogen and oxygen, people knew about water. They had the concept of water, but not the concepts of hydrogen and oxygen, and so not the concept of H2O. And so they didn’t know that water is H2O. Even after hydrogen and oxygen were discovered, someone may have thought ‘I wonder whether water is made of hydrogen and oxygen or something else’. So ‘water is H2O’ is not analytically true. On this understanding, it is conceivable, or logically possible, that water is not H2O.

But water and H2O are one and the same thing – the two concepts refer to just one thing in the world. Water is identical to H2O. Now, nothing can be what it is not. So if the property of being water and the property of being H2O are one and the same property, you can’t have ‘one’ without ‘the other’. If A is the same thing as B, then A and B can’t be separated – there is just one thing here. So while we have two concepts – WATER and H2O – there is only one property that they both pick out in the world.

What this is means is that, although it is logically possible for water not to be H2O, it is metaphysically impossible for water to be anything other than H2O. It seems that not everything that is logically possible is metaphysically possible.

Why don’t we just say that it is physically impossible for water to be anything other than H2O? This claim is certainly true, but it isn’t strong enough. If the laws of nature are contingent, then perhaps they could be different. Light could still be light but travel at a different speed, couldn’t it? The claim with water and H2O is stronger. Water wouldn’t be water if it wasn’t H2O. If the laws of nature changed, so that hydrogen and oxygen never bonded and there was no such thing as H2O, then there would be no such thing as water. There couldn’t be water, but with a different chemical composition.

We now have a sense of what metaphysical possibility is, and how it is different from both physical possibility and logical possibility. What is metaphysically possible is constrained by the real nature or identity of things. We also have an example of how something could be conceivable but not metaphysically possible, namely thinking that water is not H2O but something distinct. Another example is often used to try to make the point that what we think is distinct may not always be distinct. Suppose I believe that the Masked Man has robbed the bank. I also believe that my father has not robbed the bank. Clearly, I conceive that the Masked Man is not my father. It is logically impossible for one and the same person both to rob the bank and not to rob the bank. Does this entail that it is metaphysically possible that the Masked Man is not my father?

In one sense, we might say that the Masked Man could be anyone – nobody knows who he is. But we also rightly think that whoever the Masked Man is can’t be someone else. No one can be somebody else. I can’t be you, and you can’t be me. So if my father is not the Masked Man, it is metaphysically impossible that my father is the Masked Man. And if my father is the Masked Man, then it is metaphysically impossible that my father is not the Masked Man. (You can run the same argument with Batman and Bruce Wayne…)

Now I can conceive that my father is not the Masked Man (it is logically possible). But this doesn’t show that it is metaphysically possible that my father is not the Masked Man. If the Masked Man is my father, then it is metaphysically impossible for my father to be a different person from the Masked Man. From my conceiving that ‘two’ people are distinct, we cannot infer that it is metaphysically possible that they are distinct.

We can now apply these ideas to Descartes’ argument. Descartes argues that it is possible for the mind to exist independently of the body, because he can conceive of it existing without the body. In other words, he argues that because it is logically (conceptually) possible for the mind to exist without the body, it is also metaphysically possible. But this doesn’t follow. Perhaps unknown to him, the mind is not an ontologically independent substance, and it is metaphysically impossible for it to exist separately from the body (just as it is metaphysically impossible for water to exist without H2O).

**Reply**

However, Descartes is happy to grant that we cannot in general infer what is (metaphysically) possible from what we think. But in the case of clear and distinct ideas, the inference is justified. For example, we may rightly claim that it is impossible for a triangle to have internal angles that don’t add up to 180 degrees just because it is inconceivable that they should. Likewise, because he can clearly and distinctly conceive that mind and body are distinct substances, Descartes argues, it follows that it is possible that they are.

This provides a contrast with the Masked Man. My conceptions of my father and the Masked Man are not clear and distinct in the way that Descartes requires. It is only while we do not know who we are thinking of when we think of the Masked Man that we can think that the Masked Man could be anyone. And so, Descartes would argue, the Masked Man fallacy cannot be used as an objection to his argument.

Does this response work against the case of water and H2O as well? Were people who wondered about whether water is H2O simply not thinking clearly and distinctly? Descartes could argue that they were – our sense experience, e.g. of water, doesn’t tell us what water really is. By contrast, he holds, introspection does tell us clearly and distinctly what mind is. We fully understand what the mind is by reflecting on our own mind, and we fully understand what bodies are by reflecting on our experience of bodies. Because we have this understanding, we can know that minds are not dependent on bodies to exist, and are therefore separate substances.

What is metaphysically possible tells us nothing about the actual world

Suppose that it is metaphysically possible that the mind can exist as a distinct substance. Does it follow that the mind does exist as a distinct substance?

Let us assume, for the purposes of argument, that we conceive of mind as something that thinks and of body as something that is extended. From this, it does not follow that the mind exists as something that thinks and isn’t extended or that body exists as something that is extended and does not think. There is nothing in the initial conceptions of mind and body that oppose each other. There is no contradiction in conceiving of mind as something that is extended and thinks, or again as the thinking part of something that is extended. Likewise, there is no contradiction in conceiving of body as something that is extended, but which may, in some instances, also think. If this is right, then we can conceive of mind and body as distinct substances, or we can think of thought and extension as properties of the same substance.

Assume that whatever we can clearly and distinctly conceive is metaphysically possible. Therefore, if Descartes is right about clear and distinct ideas, it is metaphysically possible that mind and body are distinct substances. But equally, given what was just argued, it is metaphysically possible that thought and extension are two properties of a single substance. What we need to know is which option is true in the actual world. Simply knowing what is metaphysically possible does not tell us which possibility correctly describes reality. So just because it is metaphysically possible for mind and body to be separate substances doesn’t show that they are separate substances.

However, we should accept that what is metaphysically impossible does tell us something about how things actually are in the world, because what is metaphysically impossible cannot exist. If Descartes could show that it is metaphysically impossible for mind and body to be the same substance, that would show that they must be separate substances. So he could argue that we cannot clearly and distinctly conceive of mind and body as anything other than separate substances – just as we cannot clearly and distinctly conceive of a triangle not having internal angles that add up to 180 degrees. Something about our concept of mind, e.g. its indivisibility, means that we cannot conceive of it as extended; and something about our concept of body (what?) means that we cannot conceive of it as having thought. But is this right?

**2.3. Property dualism**

Property dualism is the view that, although there is just one kind of substance, physical substance, at least some mental properties are not physical properties (as type identity theory claims) nor functional properties (as functionalism claims), nor are they behavioural dispositions (as philosophical behaviourism claims). Instead, they are properties that do not supervene on physical properties in the way that physicalism claims. While mental properties are possessed by physical substances, they are a fundamentally different kind of property from physical properties.

Property dualism most often defends this claim for phenomenal properties of consciousness. Consciousness, especially the sort of consciousness involved in perception, sensation and emotion, has a ‘feel’ to it, a distinctive ‘experiential quality’. The phrase often used to try to capture this experiential quality is ‘what it is like’. There is something it is like to taste beer, to see a red rose, to feel sad. These properties can’t be reduced to physical, behavioural or functional properties. These properties, at least, are a completely new type of property. (For more on phenomenal properties, see the section ‘What do we mean by mind?’.)

**Rejecting physicalism**

Property dualism rejects physicalism, and claims that there are some mental properties that exist that are neither physical nor do they supervene on physical properties. It argues that the properties identified by physics do not form the complete fundamental nature of the universe, because in addition, there are properties of consciousness. Physics misses something fundamental. When all the physical properties of the world are finalised, this does not fix or determine the properties of consciousness the way distributing paint on a canvas determines its aesthetic properties.

Property dualists are happy to allow that there may be correlations, even natural (though not physical) laws, that connect particular physical and mental properties. So it may be a law of nature that when a creature has a certain neurological property, it has a certain conscious experience. But it is metaphysically possible for these correlations to be different, for the properties of consciousness to come apart from any physical properties with which they are correlated. Mental properties are an entirely new kind of property in the world, and do not supervene on physical properties in the way that physicalism claims.

Some property dualists argue that these mental properties have their own causal powers, which can affect physical events. This is a second way in which property dualism may reject physicalism, in rejecting physicalism’s claim that non-physical causes do not contribute to the way the physical world changes over time.

**Chalmers on explaining consciousness**

Consciousness is intimately connected to the idea of ‘subjectivity’ and undergoing experiences. In The Conscious Mind, David Chalmers argues that this aspect of consciousness, the subjective quality of experience, is very different from anything else in the world, anything else that can be investigated by science, that we think of as objective.

Chalmers begins by identifying consciousness in terms of phenomenal properties. Consciousness is not the same thing as the mind, since there can be unconscious mental states (beliefs or desires that someone isn’t aware that they have) and even unconscious processes, such as thought and even unconscious perception. Consciousness is characterized by a subjective quality of experience, so that some being is conscious if there is something it is like to be that being, and some mental state is conscious if there is something it is like to be in that mental state.

This sense of ‘consciousness’ needs to be distinguished from the ability to introspect and report on what one thinks or believes or wants. It also needs to be distinguished from the ability to focus one’s attention on something or to voluntarily control one’s behaviour. In all these senses, one is ‘conscious of’ an object (e.g. by looking at it or listening to it) or ‘conscious of’ what one is doing. These are Intentional mental states, being ‘conscious of’ something in the sense of knowing about it.

Cognitive science has investigated a great deal about the mind, including the nature of consciousness in the sense just described. It has and continues to make progress on what it is for something to be conscious in these ways, including being awake, introspection, reporting mental states, self-consciousness, attention, voluntary control and knowledge. But cognitive science has had very little to say about consciousness in the first sense, the subjective quality of experience.

Chalmers argues that we really need two concepts of mind. The first is a ‘phenomenal’ concept, where minds or mental states are characterized by the subjective quality of experience. The second is a ‘psychological’ concept, characterized by what the mind does and how we explain behaviour. The phenomenal concept deals with first-person aspects of mind, the mind as experienced by the subject; the psychological concept with third-person aspects, the mind as accounted for by others and in scientific theories. Cognitive science has said little about the mind as phenomenal because, as a scientific discipline, it deals just with those states relevant to the causation and explanation of behaviour, and not subjectivity as such.

The two concepts are complementary, but distinct. In particular, the phenomenal concept can’t be reduced to or explained in terms of the psychological concept. For example, while the brain is very complex, there is no deep mystery about the idea of brains processing information, reacting to stimuli and exhibiting complex cognitive capacities like learning, memory and language. We can offer plausible evolutionary explanations of why such functions should emerge through natural selection, and plausible physical explanations of how they occur (although lots of the details are still missing). But why does subjective conscious experience occur? If we only knew facts of physics and information processing, there would no reason to suggest that such a thing exists at all. It seems, Chalmers suggests, like a new feature of the world; it is surprising. Only our first personal experience gives us reason to think it exists.

Functionalism argues that the analysis of mental states in terms of causal functional role can be applied not just to psychological states, but to all mental states, including phenomenal ones. But, says Chalmers, while we can understand how some state could have a certain causal role, it remains mysterious why it should have phenomenal properties. Functionalism gives a good account of psychological properties, including Intentionality, but not conscious experience. After we have explained the physical and computational functioning of a conscious system, we still need to explain why this system has conscious experiences.

Just as we can talk of phenomenal and psychological concepts of mind, we can also talk of phenomenal and psychological concepts of mental states. For example, we can distinguish a phenomenal concept of pain – how it feels – from a psychological one – that it is caused by damage and leads to aversive behaviour. We don’t normally distinguish the two concepts, because the two properties usually go together. In the human mind, (phenomenal) conscious experience always also involves (psychological) cognitive processing. And so we don’t have the words for describing phenomenal qualities independent of their psychological, functional properties. We tend to pick out phenomenal properties in terms of their external qualities or causal role, e.g. we define ‘a sensation of green’ in terms of being typically caused by grass, trees, etc.. But ‘a sensation of green’ isn’t just ‘a state caused by grass, trees, etc.’. We are talking of the phenomenal quality that typically occurs when we undergo a visual experience caused by grass, trees etc. We can draw similar distinctions in our concepts of emotion, desire, and other mental states.

As long as we recognise that there are two distinct concepts here, we don’t need to argue over which is more essential to pain or colour experience or emotion… itself. But we should recognise that the co-occurrence of the two properties is not a conceptual truth. This is shown by the kind of thought experiments we looked at above – we can coherently imagine the phenomenal and the psychological properties coming apart.

Thus, we can talk of ‘psychological consciousness’ and ‘phenomenal consciousness’. Many philosophical theories and psychological studies account for psychological consciousness but not phenomenal consciousness. As with other mental concepts, phenomenal consciousness involves some psychological processing, especially ‘awareness’ – having access to some information and able to use it in controlling behaviour, e.g. give a verbal report of what one sees. But while awareness may be necessary for phenomenal consciousness, as a purely psychological phenomenon, it isn’t sufficient – it is possible to be aware of some fact without undergoing an experience with a particular subjective quality.

**The easy and hard problems**

We can also talk of two ‘mind-body problems’. The easy problem is how a physical system could have psychological properties, e.g. learning and memory. This is technical, but as we said above, it is not mysterious, since it is an account of causal roles and functions. The ‘easy’ problem involves the ‘psychological’ concept of consciousness, analysing and explaining the functions of consciousness, e.g. the facts that we can consciously control our behaviour, report on our mental states, and focus our attention. Chalmers thinks that understanding how the brain works will eventually provide the solutions. So this doesn’t threaten physicalism.

The hard problem is how a physical system could have phenomenal properties, what it is like to undergo conscious experiences. How and why are certain physical processes in the brain associated with such experiences? It is significant that the progress of cognitive science with the first problem has shed little light on the second.

Setting aside philosophical behaviourism, physicalists say that these conscious experiences just are certain physical processes or certain physical states playing a particular functional role. But, Chalmers argues in ‘Consciousness and its place in nature’, a physical account of something can only explain its physical structure and function – how something is constituted and how it works. And this, he objects, is not enough to explain phenomenal consciousness. Such explanations miss out how experiences ‘feel’, what it is like to undergo them, their subjective or first-personal aspect. There is more to phenomenal consciousness than structure and function. This thought is fundamental to arguments for property dualism.

**2.4. The ‘Philosophical Zombies’ Argument**

To understand the ‘zombie’ argument for property dualism, we first need to understand the idea of a possible world. And to do this, we need to revisit the ideas of physical, logical and metaphysical possibility (see page 15).

To summarise: what is physically possible is what is possible given the laws of nature as they are in the actual world; what is logically possible is whatever is not conceptually incoherent or self-contradictory; and we didn’t give a definition of metaphysical possibility, but we said it was constrained by the real nature or identity of things. We discussed it in terms of necessary truths that are not analytic, such as ‘water is H2O’. We can understand metaphysical possibility better by talking about ‘possible worlds’.

Let’s start by talking about true and false propositions. Propositions describe ‘states of affairs’. Propositions can be true or false. A proposition that is true describes the actual world, the way things are, a true state of affairs. A proposition that is false describes the way things are not, a false state of affairs.

However, false propositions can be necessarily false or just contingently false. A proposition that is necessarily false cannot be true – it is impossible for it to be true (either logically or metaphysically). A proposition that is only contingently false describes a state of affairs that is possible, but false, given how the world actually is. For example, ‘I was born in Kenya’ is false, but could have been true.

A contingently false proposition describes a way things could be, if they were different. We can say that in some other ‘possible world’, a contingently false proposition is true, the state of affairs it describes is part of the way that world is. In some other possible world, I was born in Kenya. A possible world is a way of talking about how things could be.

Possible worlds are distinct from one another depending on what we are supposing to be true in that world. So the possible world in which I was born in Kenya is different from the possible world in which I was born in Argentina which is different from the possible world in which I don’t exist at all.

These examples – of where I was born or even not existing – describe possible worlds that are physically possible as well as metaphysically and logically possible. There is nothing physically, metaphysically or logically impossible about the state of affairs of my being born in Kenya. We can imagine much bigger differences from the actual world without leaving physical possibility, e.g. a world in which the Earth never formed or in which evolution never gave rise to human beings.

But we can also talk about possible worlds that are physically impossible, worlds in which the laws of nature are different, e.g. in which light travels at a different speed, or that contain physically impossible things, perhaps things such as angels and ghosts. If these are genuine possible worlds – ways that a world could be – then they are worlds which are physically impossible but metaphysically possible.

As we will see, philosophers disagree on which worlds are possible worlds. It is not always easy to tell. For example, is there a possible world in which water is not H2O? In our previous discussion of metaphysical possibility, we said that ‘water is H2O’ is not an analytic truth, so it is conceivable (logically possible) that water is not H2O. But water and H2O are identical – just one thing. So it is not metaphysically possible for water to exist without being H2O. There is no possible world in which water exists, but is something other than H2O.

We can say that what is metaphysically possible is what can exist or occur as part of a possible world. Metaphysical possibility is narrower than logical possibility, since it turns out that not all conceptually coherent propositions describe how things could exist. So physical possibility concerns how things can be given the actual laws of nature; metaphysical possibility concerns how things can be in any possible world; and logical possibility concerns whether a proposition is conceptually coherent.

**Chalmers’ zombie argument**

Property dualism claims that phenomenal properties (which many property dualists claim are qualia) are not physical properties, nor do they supervene on physical properties. It rejects physicalism. But how can a property dualist show this? In ‘Consciousness and its place in nature’, David Chalmers uses the idea of possible worlds to make the argument.

According to physicalism, everything that exists is either physical or depends on what is physical. So if physicalism is true, a possible world that is an exact physical duplicate of our world (the actual world) will be an exact duplicate of our world *in all respects*. This is just the claim of supervenience, but at the level of the world. Consider: a painting that is an exact physical duplicate of another painting has all the same aesthetic properties as that painting. So a whole world that is an exact physical duplicate of another world also has all the same aesthetic properties. But what goes for aesthetic properties goes for all properties, according to physicalism. There can be no difference in, say, mental properties without a difference in physical properties. In other words, it is *metaphysically impossible*, says physicalism, for two worlds to have the same physical properties and different mental properties, because the physical properties determine the mental properties.

Therefore, if there is a possible world that is an exact physical duplicate of our world but is different in any way, e.g. it has different phenomenal properties, then physicalism is false. If two physically identical worlds have different properties of consciousness, those properties of consciousness don’t depend on physical properties. This is what Chalmers tries to show with the idea of a philosophical zombie.

**What is a philosophical zombie?**

A ‘zombie’, in the philosophical sense, is an exact physical duplicate of a person – you, for instance – but without any conscious subjective quality of experience. It therefore has identical physical properties to you, but different mental properties – it has no phenomenal consciousness.

Of course, zombies are not possible in the actual world. They are not physically possible, i.e. given the laws of our universe, we have every reason to believe that any being that has identical physical properties to you will also have consciousness.

What we are thinking about when thinking about zombies is a different possible world – a world which has all the physical properties of our world but without consciousness. We are describing a world that may be metaphysically possible.

But is such a world really (metaphysically) possible? To argue that a world with zombies is possible is to argue for property dualism. How does the argument work?

**The argument**

First, it seems that zombies are at least conceivable. I’ve just described them, and there isn’t an obvious contradiction in the idea. Second, given their conceivability, we may argue that zombies are therefore metaphysically possible. There is a possible world which has all the same physical properties as the actual world, but has no properties of consciousness.

Now, if consciousness were *identical* with physical properties, it would be impossible for a creature to have the same physical properties as you but not have consciousness. This is Leibniz’s principle of the indiscernibility of identicals. As we saw with water and H2O, if *A* is identical to *B* – if *A* is *B* – then you can’t have *A* without *B* or vice versa; they are the same thing. So if zombies are possible – if a creature could be physically identical to you but not have consciousness – then consciousness is *not* identical to any physical properties. So, if zombies are metaphysically possible, then consciousness is not identical to any physical properties. Furthermore, if zombies are metaphysically possible, consciousness doesn’t supervene on physical properties either, because you and your zombie ‘twin’ have identical physical properties, but different phenomenal properties. And so property dualism is true: phenomenal properties are neither reducible to nor supervenient upon physical properties.

P1. It is conceivable that there are zombies.

P2. If it is conceivable that there are zombies, it is metaphysically possible that there are zombies.

C1. Therefore, it is metaphysically possible that there are zombies.

P3. If it is metaphysically possible that there are zombies, then phenomenal properties of consciousness are neither physical properties nor supervene on physical properties.

C2. Therefore, phenomenal properties of consciousness are neither physical properties nor supervene on physical properties.

C3. Therefore, physicalism is false and property dualism is true.

**Responses to the zombie argument**

In this section, we consider three possible responses physicalists may make to the argument. First, they may argue that what is being proposed – in this case, a possible world that contains zombies – is not conceivable (P1 is false). Second, they may argue that although zombies are conceivable, they are not metaphysically possible (P2 – and therefore C1 – is false). Third, we may argue that even though zombies are metaphysically possible, this doesn’t tell us what consciousness is, and its relation to physical properties, in the actual world (P3 is false).

**A philosophical zombie (or zombie world) is not conceivable**

The first premise of the zombie argument claims that we can conceive of beings that have the same physical properties as us but without consciousness. Why think this is conceivable? Because when we think of physical properties, this doesn’t determine what we must think of consciousness. By contrast, when we think of the answer to 3 × 4, we must – if we are thinking clearly – think of 12. It is inconceivable that 3 × 4 is anything other than 12. Or to use an example from Descartes, it is inconceivable that the internal angles of a triangle could add up to anything other than 180 degrees. By contrast, it does not seem inconceivable that there could be a being with identical physical properties to you, but without consciousness.

The first objection to the argument is that, despite appearances, zombies are not conceivable. If we think they are conceivable, we are not thinking clearly or we lack some relevant information. It is difficult to recognise that we are not thinking clearly. But we can spell out where we are going wrong in more detail.

First, if physicalism is true, we should note that something’s physical properties determine its functional properties. So a physical duplicate of you is also a functional duplicate of you. (If physicalism is not true, then something’s functional properties could depend on its non-physical properties as well. But we cannot assume that physicalism is false, since that is what the zombie argument is trying to prove. To assume physicalism is false is to beg the question.)

Second, we need to revisit the arguments that phenomenal consciousness can be analysed in terms of physical and functional properties; there are no qualia. If we are not persuaded by this claim, it is probably because our analysis of consciousness is still underdeveloped. But if we had a complete analysis, we would see that consciousness can be completely explained in these terms. In that case, a physical, functional duplicate of you would also have consciousness.

So, once we are clear on a being’s physical properties, we can, in principle, deduce how it functions, and from this, with a complete analysis of consciousness, we can deduce whether or not it is conscious. So to imagine a being with identical physical properties to you but without consciousness is confused. It is like accepting the premises of a deductive argument but rejecting the conclusion. In conceiving of a ‘zombie’ as having identical physical properties, you conceive of it as having identical functions. But to function in certain (highly complex) ways just is to be conscious. So zombies – physically identical, but non-conscious beings – are inconceivable. (As Descartes might put it, the ideas of a zombie and of consciousness are not clear and distinct. When we make them clear and distinct, we see the contradiction in thinking that zombies are possible.)

P1. A zombie is a physical duplicate of a person with phenomenal consciousness, but without phenomenal consciousness.

P2. (If physicalism is true,) A physical duplicate is a functional duplicate.

C1. Therefore, a zombie is a physical and functional duplicate of a person, but without phenomenal consciousness.

P3. (If physicalism is true,) Phenomenal properties are physical properties realising particular functional roles.

C2. Therefore, a physical and functional duplicate of a person with phenomenal consciousness has phenomenal consciousness.

P4. A physical and functional duplicate of a person with consciousness cannot both have and lack phenomenal consciousness.

C3. Therefore, (if physicalism is true,) zombies are inconceivable.

This objection to the zombie argument depends on there being a complete physical and functional analysis of consciousness. I have inserted the phrase ‘if physicalism is true’ into two premises, (P2) and (P3), which the property dualist will contest. If there is no such analysis, because an analysis of consciousness in terms of its physical and functional properties doesn’t provide an analysis of what it is like to experience something, then it seems that this response to the zombie argument fails.

Should we accept (P2) and (P3)? To do so, we need to have good reasons to think that phenomenal properties can be understood or explained either in terms of physical structure or in terms of functions. But there is nothing in our phenomenal concept and experience of consciousness that supports the claim. And so we can conceive of that same physical thing either with or without phenomenal consciousness.

The debate looks like a stalemate. On the one hand, the zombie argument mustn’t assume that physicalism is false, since it is trying to show that physicalism is false. On the other hand, the response seems to assume that physicalism can give a complete analysis of our concept of consciousness.

Many philosophers have concluded that we should grant that zombies are conceivable, and focus the discussion on whether they are metaphysically possible. That takes us to our second response.

**What is conceivable may not be metaphysically possible**

The second response targets the second premise of the zombie argument. Although zombies are conceivable, they aren’t in fact metaphysically possible. What we are able to conceive is not always a reliable guide to what is possible.

**Identity and metaphysical possibility**

To understand this, let us return for once again the example of water and H2O. As we saw, the two concepts WATER and H2O are distinct, and it is not an analytic truth that water is H2O. So it is conceivable (even if false) that water is not H2O.

Given this, it is easy to think that water could have been different, i.e. in some possible world, water is not H2O. However, given that water is H2O, it’s not metaphysically possible that water isn’t H2O. This was an important claim about identity first made by Saul Kripke in *Naming and Necessity*. It’s not possible for A to be B and for it not to be B. So if A is identical to B – if A is B – then A is B in every possible world. Because water is H2O, it is H2O in every possible world.

It is possible that the water in the oceans could have been fresh, not salty. Or in other words, in another possible world, the water in the oceans is fresh, not salty. The fact that oceans are salty is a contingent property of water in our world. It isn’t what makes water what it is. Or again, the fact that water falls as rain is a contingent property of water. If it never rained, this wouldn’t change what water is. So in another possible world, water never falls as rain.

But turn now to the question of what makes water what it is? What is the essential property of water? The answer: its chemical composition, H2O. Now, what makes water what it is is not a property that water can lack in some possible world. A world without H2O is a world without water, because water just is H2O.

Suppose there is another possible world in which a transparent, odourless liquid falls as rain, fills the oceans, freezes and evaporates, etc. but isn’t H2O. Is this liquid water? No, says Kripke. It is something just like water, in that it has many of the contingent properties of water. But it isn’t water, because it isn’t H2O.

Kripke concluded that identity claims – ‘A is identical to B’ – are necessarily true, if true at all. They are true in all possible worlds.

We said that we can conceive of water not being H2O. But we have argued that it isn’t possible that water is not H2O. This shows that we cannot always infer metaphysical possibility from conceivability.

**The response to the zombie argument**

We can now apply the point to zombies. The fact that we can conceive of zombies doesn’t show that zombies are metaphysically possible. If phenomenal properties just are certain physical and/or functional properties, then it isn’t possible for zombies to exist (even if they are conceivable). Given the physical properties we have, if physicalism is true, it just isn’t possible for a being with the same physical properties not to have consciousness as well. If physicalism is true, then when we think of phenomenal consciousness and, say, certain neurological or functional properties, we are thinking of one and the same property in two different ways, using two different concepts.

This response doesn’t have to claim that phenomenal properties are physical properties, that physicalism is true. It only has to claim that the zombie argument cannot show that physicalism is false. The premise that zombies are metaphysically possible cannot be defended without assuming that phenomenal properties are not, unknown to us, physical properties.

**A disanalogy?**

This second objection to the zombie argument relies on an analogy between phenomenal consciousness and scientific identities, such as water and H2O or life and chemical processes. Property dualists can argue that this analogy doesn’t work.

Something isn’t water if it isn’t H2O, because H2O is the ‘essence’ of water. The concept WATER is a concept of something that has a particular structure and causal role, which science can then discover. Water is precisely the kind of thing that could be – and is! – identical with a chemical property. This is why you can’t have water without H2O or H2O without water.

By contrast, say property dualists, the essence of phenomenal properties is what it is like to experience them. The essence of pain – what makes pain pain – is how pain feels. Its essence isn’t some physical or functional property. The essence of a physical property is its physical structure or composition; the essence of a functional property is what causes it and what it causes. In arguing that neuroscience can tell us what consciousness ‘really is’, physicalists are assuming that the essence of consciousness, like the essence of water, is something physical. But this is a mistake. Consciousness is essentially first-personal, i.e. what it is like for the person. The concept of consciousness is not the concept of something that has a particular physical structure or set of causal relations that science can then discover. So consciousness is not essentially a set of brain properties described by the neuroscientist.

If this is correct, then the correlation between brain properties and consciousness in the actual world is contingent. As it happens, certain brain processes give rise to consciousness. But you could have the brain processes without consciousness and vice-versa. It is only the essential properties of something that can’t change in different possible worlds, the contingent properties can. The same physical processes that are correlated with consciousness in the actual world may not be correlated with consciousness in another possible world. Because phenomenal properties have a different essence from physical and functional properties, each can exist without the other. So zombies are possible.

What is metaphysically possible tells us nothing about the actual world

A third objection to the zombie argument targets the inference from the claim that zombies are possible to the conclusion that property dualism is true. The zombie argument shows, at best, that in another possible world, physical properties and phenomenal properties are distinct. But why does this entail in the actual world that they are distinct? Couldn’t it be the case that physicalism is true in the actual world, but property dualism is true in a different possible world? Or in other words, the zombie argument only shows that property dualism is possible; it doesn’t show that property dualism is true.

We can reply that this objection makes two mistakes. First, the objection misunderstands identity. It suggests that phenomenal properties could be physical properties in this world but not in another possible world. But this isn’t possible. Nothing can be something else. I can’t not be me in another possible world. If ‘I’ were not me, but you, say, then that person is not me. In any possible world, the only person I can be is me. Likewise, water can’t be something other than water. Since water is H2O, it can’t be something else in another possible world.

The same goes for phenomenal properties. If phenomenal properties are physical properties in this world, then they are physical properties in every possible world. And if they are not physical properties in another possible world, then they are not physical properties in any possible world, including the actual world. When it comes to identity, possibility does tell us about reality.

Second, if the objection is intended to defend physicalism, it misunderstands what physicalism claims. Physicalism claims that what exists is either physical or supervenes upon what is physical. We need to be clear about supervenience. For example, we want to say that if the physical properties of two paintings are identical, then the aesthetic properties cannot be different. It is not strong enough to simply say that they aren’t different, since that would allow that the physical properties don’t ‘fix’ the aesthetic properties.

What we said about the aesthetic properties applies to properties of consciousness as well, and what applies to paintings is true of whole worlds. According to physicalism, once the physical properties of a world are finalised, then there is no further work to be done to ‘add’ consciousness. It is already part of the world. Phenomenal properties cannot differ independently of physical properties. So physicalism is a claim about what is metaphysically possible.

The zombie argument attacks this claim. It argues that there can be two worlds that are physically identical but with different phenomenal properties. Once the physical properties of a world are finalised, then there is still further work to be done to ‘add’ consciousness. Thinking about possibility does, in this case, tell us about reality.

**Patricia Churchland on thought experiments**

In *Brainwise*, Patricia Churchland is sceptical about the use of appealing to conceivability or metaphysical possibility to discover the nature of the world. In imagining a zombie, we are imagining a being with a brain just like ours. But in imagining that it doesn’t have phenomenal consciousness, we are imagining that if we knew everything about neuroscience, we still wouldn’t have explained or understood consciousness. But all this imagining is really a reflection of our own epistemic limitations and the fact that neuroscience just isn’t very developed yet. The thought experiment of zombies don’t tell us anything significant about the nature of consciousness. Property dualists are mistaken in trying to get conclusions about what exists out of epistemic premises.

Suppose someone (perhaps 200 years ago) said ‘I just can’t imagine how living things could really be composed of dead molecules – how can life arise out of the interactions of things that are not alive?’. Or again, suppose someone proposed the thought experiment of ‘deadbies’, creatures who are physically identical to us, but aren’t alive. They claim to be able to imagine such creatures. None of this would persuade us to think again about vitalism and the existence of a special, non-reducible ‘life force’. On the current biological theory of what it is to be alive, deadbies are impossible and life really is just the highly complex interactions of molecules.

Similarly, we shouldn’t be persuaded by the property dualists’ appeal to zombies. First, from what is conceivable we cannot infer anything about the nature of how things are. Our grammar – our concepts as they are now – are not necessarily a good guide to how things really are. We change our concepts as we discover more about the world. Second, the same goes for thought experiments about what is metaphysically possible. What things really are is what they are in the actual world. We discovered what water is through scientific investigation. Similarly, the right way to think about consciousness is through scientific investigation, and we shouldn’t let considerations about concepts determine in advance what scientific investigation may or may not discover. For example, contemporary biology argues that genes are DNA. Should we object this reductive explanation because in another possible world, genes – understood as the units of heredity – might not be DNA? No. What genes are is what genes are in the actual world. And all we need to make this claim is an empirical identity, supported by scientific explanation. The same goes for consciousness.

Churchland is arguing that philosophy simply can’t do metaphysics in this way, using thought experiments and possible worlds to discover what something is. Philosophical ‘speculation’ must give way to experimental science.

**2.5. The knowledge argument**

In ‘Epiphenomenal qualia’, Frank Jackson defends property dualism on the basis of his ‘knowledge argument’. He describes the following scenario. Suppose there is a neuroscientist, Mary, who has lived all her life in a room in which everything is black and white. She has never seen any colour other than black, white and shades of grey. However, she has specialised in the science of vision, and through textbooks and black-and-white TV, she has come to know every physical fact there is to know about colour vision – everything about the properties of light, everything about the eye, everything about the nerves and the brain related to vision. So, Mary knows all the physical information there is to know about what happens when we see a ripe tomato. She is then let out of the black-and-white room, and comes to see something red for the first time. Does she learn something new?

Jackson claims that ‘it seems just obvious’ that she will. She will learn about what it is like to see the colour red. And so she learns something new about our visual experience of the world. However, we said that she knew all the physical facts while she was in the room. So not all the facts are physical facts. It is possible to know all about the physical properties of the brain involved in having an experience and yet not know about the qualia.

P1. Mary knows all the physical facts about seeing colours before being released from her black-and-white room.

P2. On being released, she learns new facts about seeing colours.

C1. Therefore, not all facts are physical facts, e.g. some facts about colours are not.

C2. Therefore, phenomenal properties are non-physical and physicalism is false.

By ‘all the physical facts’, Jackson means not only what we already know about physics and neurophysiology. Mary knows all the physical facts as discovered by a completed physics and neuroscience. Furthermore, she has worked out all the causal and functional facts that are entailed by these facts. Because physicalism claims that the world is entirely physical (if we include causal and functional properties), it must claim that to have complete physical knowledge is to have complete knowledge. But no amount of physical information can enable Mary to know what it is like to see a ripe tomato.

**Responses to the knowledge argument**

Physicalist responses to Jackson’s argument point out that there is more than one meaning of ‘to know’, more than one kind of knowledge. We can and should accept that Mary gains new knowledge when she sees red for the first time. But this doesn’t mean that she gains knowledge of some new fact. We will look at three different responses offering alternative accounts of just what Mary learns.

1. **Mary does not gain new propositional knowledge, but does gain ability knowledge**

The first response argues that instead of gaining knowledge of a fact, described by a proposition (e.g. ‘that red looks like this’), Mary gains know-how – the knowledge involved in certain abilities. For instance, to see red for the first time is to gain the ability to know how to imagine or recognise red. So Jackson hasn’t shown that there are any facts that are not physical facts.

We can challenge this objection as follows. Suppose that seeing red gives us these new abilities. Are such abilities all that is involved in knowing what it is like to see red? Suppose Mary wonders whether what it is like for others to see red is the same as what it is like for her. She isn’t wondering about her abilities to imagine and recognise red. She is wondering about the truth of a proposition. So when Mary first learns what it is like to see red, she does gain knowledge of a new fact.

Is the objection even right to think that knowing what it is like to see red involves knowing how to imagine red? Suppose there is someone who (for whatever reason) has no ability to imagine seeing red. Now suppose this person looks attentively at something red. While they look at red, they know what is it like to see red. And yet they cannot imagine seeing red. This shows that the ability to imagine is not necessary for knowing what it is like to see red. Now suppose someone else has the most amazing ability to imagine seeing colours. They are told that there is a shade of red, e.g. burgundy, that is between plum red and tomato red. They are now able to imagine burgundy, but as long as they don’t actually imagine burgundy, they still don’t know what it is like to see burgundy. This shows that the ability to imagine a colour is not sufficient to know what it is like to see it. (We can make similar arguments for recognising colours.)

If the ability to imagine seeing red is neither necessary nor sufficient for knowing what it is like to see red, then when Mary comes to know what it is like to see red, she learns more than simply knowing how to imagine seeing red. The response fails to show that Mary does not learn a new fact. It fails to show that the knowledge argument is mistaken.

1. **Mary does not gain new propositional knowledge, but does gain acquaintance knowledge**

A second response to Jackson’s argument argues that Mary gains a different kind of knowledge again, not propositional knowledge (knowing that), but not ability knowledge (knowing how) either. Instead, she gains ‘acquaintance knowledge’ – knowledge given by direct awareness of something in experience, e.g. a person, a place, or one’s own thoughts and feelings. To see red is a direct apprehension of red, as contrasted with descriptions of seeing red. How does the objection work?

Suppose that what it is like to see red is a physical property of the visual experience, which itself is a physical process. In other words, the phenomenal property of what it is like to see red is some property of the brain (type identity). Mary can then know all about this physical property, about what it is, when it occurs, and so on, before she leaves the room. However, she is not acquainted with the property – she doesn’t have direct knowledge of it because her brain has never itself had this property. When she sees red, this property occurs in her brain and she becomes acquainted with it. She gains new knowledge, but she hasn’t learned any new fact. She already knew all about this property before she left the room. (Compare: a friend describes someone you have never met. When you first meet the person and become acquainted with them, you think of them in a new way. But the person you meet was someone you already knew about.)

In *Brainwise*, Patricia Churchland puts the two responses together. Knowing the neuroscience won’t help you experience or identify phenomenal properties in consciousness. For this, the theory needs to be true of your brain, i.e. your brain needs to undergo the processes that the theory describes as constituting colour experience. This fact doesn’t mean that there is something that the theory misses out. When Mary’s brain actually undergoes the processes that she knows all about, then she will be acquainted with colour and gain abilities of recognition etc. But that is all the colour experience is. Nothing in addition to the physical processes is needed or occurs.

There are two possible responses to this objection. First, we can argue that acquaintance knowledge involves propositional knowledge. What it is to be acquainted with red is to know that seeing red is like this (having the experience). Becoming acquainted with red involves learning some new fact. So Mary does learn a new, and therefore non-physical, fact when she becomes acquainted with red. So what it is like to experience red can’t simply be a physical property of the brain.

Second, we can argue that the objection misunderstands the argument. The knowledge argument isn’t about Mary’s experience. The argument is that Mary didn’t know everything about other people’s experiences before she left the room, even though she knew everything physical about their experiences. Mary doesn’t know what it is like for anyone to experience red. This is a fact about experiences that Mary doesn’t know. When Mary leaves the room, she realises how impoverished her conception of people’s colour experiences has been. So there are facts about other people’s experiences of seeing red that Mary learns.

P1. Mary (before her release) knows everything physical there is to know about other people when they see colour.

P2. Mary (before her release) does not know everything there is to know about other people when they see colour (because she learns something about them on her release).

C1. Therefore, there are truths about other people (and herself) when they see colour which escape the physicalist story.

C2. Therefore, phenomenal properties are non-physical and physicalism is false.

1. **Mary gains new propositional knowledge, but this is knowledge of physical facts that she already knew in a different way.**

A third response to Jackson’s argument distinguishes between two ways we might talk about ‘facts’ on the basis of the distinction between concepts and properties.

Suppose I know that there is water in the glass. Is that the same as knowing that there is H2O in the glass? No – because someone may know one of these truths without knowing the other. Someone can have the concept WATER without having the concept H2O. Or again, someone may have both concepts, but not know that water and H2O are the same thing. So we can say that to know that there is water in the glass and to know that there is H2O in the glass is to know two different facts. In this sense of ‘fact’, we count facts in terms of concepts.

However, in another sense of ‘fact’, the fact that there is water in the glass just is the fact that there is H2O in the glass, because water and H2O are identical – one thing. Both of these claims are made true by just one state of affairs in the world. In this sense of ‘fact’, we count facts in terms of how the world is, not how we think about it. Another way of expressing this is to say that the property of being water and the property of being H2O are one and the same property. We will use ‘fact’ in this sense from now on.

We can now apply this to the knowledge argument. Before leaving the room, Mary has a concept of red in physical terms – wavelengths of light, neurons firing, and so on. Call this the ‘physical’ or again a ‘theoretical’ concept of red, REDTH. Or again, using Chalmers’ distinction between psychological and phenomenal concepts, Mary knows what it is to see red in the psychological sense of ‘seeing red’. We can contrast this with a ‘phenomenal’ concept of red, REDPH. A phenomenal concept of something is the concept by which you recognise something when you experience or perceive it. So we gain the phenomenal concept REDPH by seeing red. Before she leaves the room, Mary doesn’t know what it is to see red in the phenomenal sense.

When Mary comes out the room and sees red, she acquires the phenomenal concept REDPH for the first time. She is now able to think about red in a new way, in terms of what it is like to see red. She couldn’t know what it is like to see red before because she didn’t have the phenomenal concept. But, we can claim, the phenomenal concept REDPH is a concept of the same thing that her theoretical concept REDTH is a concept of – they are two different concepts of a physical property of the brain (like WATER and H2O are two concepts of the same physical substance). Mary gains new propositional knowledge about seeing red in one sense (because she gains a new concept) but her new knowledge is about a property that she already knew about under a different concept. The theoretical concept REDTH and the phenomenal concept REDPH are two concepts that refer to the same property.

Let us accept that the knowledge argument shows that there are different ways of thinking about physical things, some of which depend on experiencing, rather than describing. To know what it is like to see red, you need to have the phenomenal concept REDPH, and this you can only gain from experience. So Mary gains knowledge of a new fact, in the sense of fact that relates to concepts.

However, physicalism and property dualism are claims about what exists. They are claims about properties, not about concepts. The knowledge argument does not show that Mary gains knowledge of a new property. It doesn’t show that Mary learns about something in the world that she didn’t know about before. And so it doesn’t show that what it is like to see red cannot be a physical property. So the argument fails to show that there are any non-physical properties. So it fails to show that physicalism is false.

**2.6. The problem of other minds**

The problem of other minds is the question of how we can know that there are minds other than our own. We each experience our own minds directly, from ‘within’. We can each apprehend our sensations and emotions in a way that is ‘felt’. We can know what we want or believe through introspection. But our knowledge of other people’s minds is very different, it seems. We cannot experience other people’s mental states. It seems that all we have to go on is other people’s behaviour, what is expressed through their bodies.

While this is a general philosophical puzzle, it raises a special challenge for substance dualism, which holds that minds are distinct substances from bodies – they are not bodies, they are not parts of bodies, and because they are substances, they are not properties of bodies either. Cartesian dualism – the form of substance dualism defended by Descartes – also claims that minds do not depend on bodies in order to exist, i.e. minds can exist separated from any body. People who believe that the mind is the soul, and the soul can continue to exist without a body after death, are usually substance dualists. But if minds and bodies are entirely independent, then how can I infer from seeing a body that there is a mind ‘attached’? The two things exist independently of one another. So other ‘people’ – other bodies – could all be machines, programmed to behave as they do, but with no minds. How can I know otherwise?

**The argument from analogy**

The argument from analogy claims that we can use the behaviour of other people to infer that they have minds too.

P1. I have a mind.

P2. I know from experience that my mental states cause my behaviour.

P3. Other people have bodies similar to mine and behave similarly to me in similar situations.

C1. Therefore, by analogy, their behaviour has the same type of cause as my behaviour, namely mental states.

C2. Therefore, other people have minds.

The argument is perhaps the ‘common-sense’ position on how to solve the problem of other minds. But we can object to its use of induction. The conclusion that other people have minds is based on a single case – mine. This is like saying ‘that dog has three legs; therefore, all dogs have three legs’. You can’t generalise from one case, because it could be a special case. Perhaps I am the only person to have a mind. And we can’t get around this by first checking that other people have minds to show that I am not a special case!

However, instead of talking about the causal relation in the single case of my behaviour and my mind, we can formulate the argument to cite many instances of behaviour which we know to have a mental cause.

P1. This behaviour has a mental cause.

P2. That behaviour has a mental cause.

P3. That third behaviour has a mental cause.

P4. Etc.

C1. Therefore, many behaviours have a mental cause (I know this from my own experience).

P5. Other people exhibit the same types of behaviour as cited above.

C2. Therefore, those behaviours also have mental causes.

C3. Therefore, other people have minds.

However, this faces two objections. First, although many behaviours of which I have experience have mental causes, not all of them do. Sometimes I do something without being aware of a mental cause. So while (C1) is correct, it isn’t strong enough to support the claim that the behaviour of other people also has mental causes – perhaps, like some of my own behaviour, it does not. Second, the argument relies on the contentious claim that similar effects (behaviour) have similar causes (mental states). But sometimes similar effects can have different causes. Perhaps those instances of other people’s behaviour that are similar to my behaviour have different (non-mental) causes.

**The existence of other minds is the best hypothesis**

Rather than inferring from one’s own case to other minds, we may employ a standard form of theoretical scientific reasoning, inference to the best explanation. This argument doesn’t appeal to the first-personal experience of having a mind nor does it draw an analogy between my behaviour and that of other people. Instead, the question is entirely third-personal. Why do human beings behave as they do? What hypothesis best explains people’s behaviour in general? The claim is that the best explanation is that people have minds, and that their mental states cause them to behave as they do. And if people in general have minds, then obviously people other than me have minds.

Why think that the best hypothesis for explaining human behaviour is that people have minds that cause their behaviour? In particular, why think that it is a better hypothesis than the claim that people are machines without minds?

One way philosophers have developed the argument is to analyse mental states as the ‘inner’ states of an organism that respond to the environment and cause behaviour – this is what mental states are. Pain makes you respond quickly to prevent further damage; desire makes you pursue something you need; belief gives you information you need in order to pursue desires; and so on. The theory that there are such ‘inner’ states that cause behaviour is then said to be the best explanation of behaviour. (A substance dualist can then argue that these ‘inner’ states are states of a distinct mental substance, rather than, say, states of the brain.)

This line of argument faces three challenges. First, it depends on this theory being the right account of what mental states are.

Second, if we understand the mind in terms of its causal relations to behaviour, then we need to solve the problem of how the mind can cause physical events. But substance dualism has difficulties in explaining mental causation. However, the dualist can respond that to solve the problem of other minds, we only need the claim that behaviour is caused by mental states, not an explanation of how.

Third, we can object that the belief that people have minds is not a hypothesis, nor do we infer, on the basis of evidence, that they have minds. Consider: have you ever seriously wondered whether people have minds and then used their behaviour as evidence that they do? This whole way of understanding the way we think about minds is mistaken.

**Avramides on Descartes’ solution**

The problem of other minds seems particularly challenging to substance dualism, because it claims that mind and body are completely separate things. So a human body, it seems, really could be just a machine, programmed to behave in certain ways but with no thought or consciousness, because these properties belong to something entirely different to a body. So what did Descartes say about how we know that other people have minds? Many philosophers have thought that he offers a version of the argument from analogy, but in *Other Minds*, Anita Avramides argues that this is a misinterpretation.

At the end of *Meditation* II, Descartes says ‘if I look out of a window and see men crossing the square… I say that I see the men themselves… Yet do I see more than hats and coats which could conceal robots? I judge that they are men’. This can sound like he is entertaining the idea that what we ‘see’ when looking at other people could either be men, i.e. with minds, or machines, i.e. without minds. But we infer – ‘judge’ – that they are men.

But, argues Avramides, there are two objections to understanding Descartes in this way. First, the argument from analogy uses two separate observations and two separate judgments, viz. that others behave as I do, and so (by analogy) others must have a mind as I do. By contrast, Descartes suggests there is just one judgment in observing another mind.

Second, the context for this passage in Descartes is the relation between perception and judgment. Using an example of wax that undergoes changes when heated, Descartes has argued that our sense experience doesn’t give us knowledge of what physical objects are. The sensory qualities of the wax change, but we judge that the wax remains the same thing. It is our judgment, not our perception, that gives us knowledge that the wax exists. Although we say we ‘see’ the wax (through vision), in fact we judge (through understanding) that it is present from what we see. Similarly, we ordinarily say that we see men, but we really see hats and coats, and from this, judge that what we see are men. Descartes does not say we see the body of a man and then judge that this body has a mind. And unlike the argument from analogy, he does not proceed from his own case – that he has a mind is not a premise in his argument. He is simply not discussing the problem of other minds.

In fact, Descartes never discusses the problem of other minds as such. However, we can find his solution to the problem by looking at what he says about people and animals. Descartes believes that, unlike people, animals are machines, without minds. All bodies, insofar as they are bodies alone, are machines. Their functions and their behaviour can be seen to follow from the arrangement of their parts. This can seem to raise the problem of other minds in a very acute form: How do we know that other people aren’t simply like other animals, without minds?

Descartes’ answer is that there are two differences between animals (without minds) and people (with minds). First, people use language, while animals don’t use language. It is inconceivable for a machine to answer questions meaningfully, Descartes claims, so our use of language shows that we have a mind while they do not. Second, people are capable of highly adaptable behaviour, and while some animals can show adaptable behaviour is some situations, they lack adaptability in others. If they had genuine intelligence (mind), there would be no such asymmetry.

So, a person is both mind and body, united. What distinguishes a person is language and adaptable behaviour. When we encounter others, we observe not just their hats and coats, but their use of language and adaptable behaviour. On this basis, we judge that this other is a real man (i.e. with a mind), not a machine.

**2.6. Property dualism: objections**

**Interactionist property dualism**

Nothing seems more obvious than that the mind and the body interact with each other, e.g. I decide to phone a friend and move my body to do so. Cartesian substance dualism – the claim that mind and body are distinct substances – has great difficulty explaining how this can be true. How is it that a mental substance, which is not in space and has no physical force, can affect a physical substance, which is in space and moved by physical forces?

In ‘Consciousness and its place in nature’, David Chalmers argues that property dualism doesn’t face this particular issue, because mental properties are properties of physical objects. The claim is simply that these mental properties make a difference to how the physical world changes. For instance, having a painful experience makes a difference to what I do next, e.g. jumping up and down – my bodily movements are caused by my being in pain.

How? Can interactionist property dualism provide any details of how mental properties would cause physical effects? It seems not. In reply, Chalmers notes that this is true of any fundamental causal relationship. For instance, for many years, physicists had no account of how gravity works. Then Einstein suggested that it was the result of mass bending space. But at present, we have no account of how mass bends space. But this is no objection to accepting the claim that mass does bend space. Property dualism claims that mental properties are fundamental in the same sense as fundamental physical properties. There is no further explanation in other terms available. But there is no special problem of mental causation here.

(Chalmers notes that there are interpretations of quantum mechanics that actually suggest that consciousness plays a causal role in physical events. If you are interested in physics: this is the interpretation that maintains that conscious observation of a quantum system collapses its superposed state to a determinate state.)

We can object, however, that the claim that mental properties cause physical effects is incompatible with neuroscience. Current science indicates that movements of the body are caused by physical events in the brain. So, if mental properties move the body, they do so by changing what happens in the brain. We may object that we have no evidence of mental properties changing what happens in the brain.

That is true, but we have no evidence that the claim is false either. This is because, while neuroscience is making good progress, we still have no clear account of the very complicated causation involved in something like making a choice. But we may think that neuroscience could discover the complete story in time. If interactionist property dualism is true, then it seems that what it must discover is that some events in the brain have no physical cause, because they are caused by mental properties.

**Epiphenomenalist property dualism**

If the knowledge and zombie arguments work, then property dualism is true, it seems. On the other hand, Chalmers argues, the claims of physicalism that physical laws govern all events in space-time and that every physical event has a sufficient physical cause seem appealing in light of the success of empirical science. Epiphenomenal property dualism allows both sets of claims to be true. Some mental properties are neither physical nor supervenient on physical properties, but they don’t make any causal difference to the world. Physicalism is right about causation, it just isn’t right about what exists.

**The phenomenology of our mental life**

We can object, however, that epiphenomenalism is very counter-intuitive. It is part of our experience of having mental states that our mental states, e.g. feeling pain or wanting chocolate or believing that Paris is the capital of France, cause other mental and physical states and events. Most obviously, mental states can cause our behaviour, such as wincing or going to the food cupboard, and they can be part of a causal mental process, such as thinking about how to get to Paris. The ‘phenomenology of our mental life’ involves experience of such causal connections, doesn’t it?

The epiphenomenalist property dualist replies first, that it is only those mental properties that they are dualist about that are epiphenomenal. So, for Chalmers, it is only phenomenal properties of consciousness that are epiphenomenal. We can say that beliefs and desires have causal powers, since we can analyse these states in terms of physical properties and functions. Nevertheless, that the feeling of pain or longing of love is epiphenomenal is still counter-intuitive. So, second, the epiphenomenalist property dualist offers an alternative explanation of why it seems this way to us, even though such mental properties never cause anything.

The physical process in the brain with which phenomenal properties are correlated causes both the phenomenal property, e.g. the painful experience, and the behaviour which we think is caused by the phenomenal property, e.g. jumping up and down. So the experience and the behaviour are correlated because they are both effects of the same cause. It is this correlation that makes us think that the experience causes the behaviour. But it doesn’t.

This may be counter-intuitive, but that is not sufficient reason to reject epiphenomenalism.

**Natural selection**

The property dualist believes that mental properties are properties of physical objects, namely certain living creatures. Suppose that Darwin’s theory of evolution by natural selection is true. According to this theory, millions of genetic alterations randomly take place. Most disappear without a trace. But some that coincidentally help a creature to survive and reproduce slowly spread. That creature and its descendants reproduce more than others without those traits, so more and more creatures end up with them. The features enables the creature to reproduce more, so its descendants also have that feature and they reproduce more and so on.

So, according to the theory of evolution, the traits that evolve over time are ones that causally contribute to the survival and reproduction of the creature. We can assume that mental properties, including qualia, evolved. But how, if they make no difference to what creatures do and so whether they survive and reproduce? Epiphenomenalism conflicts with our best account of the origin of consciousness.

In ‘Epiphenomenal qualia’, Frank Jackson considers this objection and replies that natural selection is more complicated than just described. In fact, there are lots of traits that have evolved that don’t contribute to survival or reproduction, but are instead by-products of traits that do contribute. For instance, polar bears have thick, warm coats which help them survive in the Arctic. A thick coat is a heavy coat. But having a heavy coat doesn’t contribute to the polar bear’s survival, because it makes the bear slower. However, it is better to have a thick, warm and heavy coat than a thin, cool and light coat. Having a heavy coat is a by-product of having a thick, warm coat, and having a thick, warm coat contributes to survival.

Likewise, there are brain processes that make a difference to how a creature behaves and which are very conducive to survival. Consciousness, according to epiphenomenalism, is simply a by-product of these brain processes. It just happens to be a fundamental law of nature that these physical properties are correlated with certain properties of consciousness.

We can object that this response presents us with a very divided picture of the world. Consciousness sits entirely outside the rest of the natural world, and has no effect on it.

Jackson accepts this: we shouldn’t expect to understand the world. Our abilities to understand the world themselves relate to survival. As a result of evolution, we are equipped to learn about and understand what we need to know in order to survive. Consciousness doesn’t make any difference to this, so it is no surprise that we can’t understand it well.

**Introspective self-knowledge**

Epiphenomenalism makes it hard to understand how we have knowledge of our own mental states. How do I know that I am in pain when I am? The obvious answer is that my belief that I am in pain is caused by my pain itself. I can tell that I am in pain just from introspection. But if epiphenomenalism is true, pain doesn’t cause anything, even my belief that I am in pain. This threatens a natural account of our knowledge of our mental states. If my thoughts and feelings don’t cause my beliefs about my mind, then I could have those beliefs whatever my mental states, just as long as the causes of my beliefs (whatever they are – perhaps brain processes, perhaps God) operate in the same way. In other words, whatever causes me to belief that I am in pain could cause me to have this belief even when I am not in pain. And so my beliefs about my mind, therefore, are unjustified and unreliable. So I can’t know my own mind.

Epiphenomenalists can reply that knowledge of something doesn’t always require that thing to cause one’s belief. I can know that I am in pain without the painful experience causing this knowledge. For instance, suppose the brain state that my belief that I am in pain is also the same brain state that causes my pain. In this case, I wouldn’t, under normal circumstances, have the belief that I am in pain unless I was in pain – the same brain state causes both. So even though my belief that I am in pain isn’t caused by the painful experience, I can know that I am in pain because my belief is caused by a reliable mechanism.

Chalmers gives a different response. Knowledge of my experiences is knowledge by acquaintance. I am directly aware of my experiences, but this is not a causal relation. My belief that I am in pain is partly constituted, not caused, by this direct awareness. My being in pain makes my belief the belief that it is. So my knowledge that I am in pain depends on my being in pain, but is not caused by it.

**Property dualism makes a ‘category mistake’**

Gilbert Ryle calls substance dualism ‘the dogma of the Ghost in the Machine’. The mistake that it makes, he argues, is a ‘category mistake’. What does that mean? Suppose someone is shown around Oxford University – they see the colleges, the buildings with the different faculties and departments, the administrative buildings. But then they ask, ‘I’ve seen the colleges, the faculties, the administration. But where is the university?’ They have misunderstood the concept of ‘university’, thinking that the university is another thing, alongside the colleges, faculties and administration. The person has made a category mistake. The university is not like this; it is how everything that the person has seen is organised.

Concepts belong to different logical categories – different ways in which it makes sense to use a concept. A category mistake is to treat a concept as belonging to a different logical category from the one it actually belongs to. According to Ryle, substance dualism makes the category mistake of thinking that the mind is like the body – another ‘thing’, a distinct, complex, organised unit subject to distinct relations of cause and effect. The mistake is to think that physical and mental concepts operate in the same way, in the same logical framework of ‘things’ and ‘causes’, ‘substances’ and ‘properties’.

We can apply his objection just as easily to property dualism. While property dualism doesn’t claim that the mind is a distinct substance, it does think of mental properties – or at least phenomenal properties of consciousness – as part of the same metaphysical framework as physical and functional properties, only not physical.

Ryle would argue that the concept of phenomenal properties (let alone qualia) misunderstands our talk of sensations, feelings, images, and so on. These are not each a ‘something’ that has peculiar properties of ‘what it is like’. The whole metaphysical picture here is wrong.

So how should we understand our talk about conscious experiences? On Ryle’s behalf, we could argue that when we express our experiences, we use words that derive their meaning from describing physical objects. To say ‘what it is like’ to see red is simply to describe what we see when attending to the colour of a red object, or if it is not in front of us, we give a report of our memory of seeing it. The redness that we experience is the redness of the rose, not a property of our experience of it.

People don’t normally talk about ‘sensations’ or ‘what it is like’ in the sense of qualia in everyday language, before being exposed to some theory. If you ask someone ‘what it is like’ to see a rose, they will usually respond evaluatively, e.g. ‘it’s wonderful’ or ‘it’s calming’. Of course, experiences differ from each other. But this isn’t because what each experience ‘is like’ differs. We can express the difference between what experiences ‘are like’ in terms of what the experience is of (red roses look different from yellow roses – this is a difference between roses, not between experiences of roses), and how we evaluate experiences, e.g. whether we enjoy one and find another boring. The property dualist has misunderstood our mental concepts.

For example, in response to the knowledge argument, what we should say is this. In knowing all the physical facts, Mary can’t yet understand our normal way of talking about experiences. She has no experiences of coloured objects that she can express and report, and as a result, she has only a limited understanding of our discussions of them. But none of this has to do with knowledge of facts, either facts about some ‘inner’ conscious experience or facts about the brain. To think otherwise is a category mistake.

**Substance dualism: can mind and body interact?**

Substance dualism claims that both minds and bodies – physical objects – exist. It is common in contemporary philosophy of mind to assume that bodies exist, and we shall share that assumption. Substance dualism is controversial, therefore, in claiming that the mind is an ontologically distinct substance.

Substance dualism holds that there are two fundamentally different types of substances. In traditional dualism, these two types of substances are physical substances (‘bodies’, physical objects) and mental substances (minds). Minds are distinct from bodies – they are not bodies, they are not parts of bodies, and because they are substances, they are not properties of bodies either. Cartesian dualism – the form of substance dualism defended by Descartes – also claims that minds do not depend on bodies in order to exist, i.e. minds can exist separated from any body. People who believe that the mind is the soul, and the soul can continue to exist without a body after death, are usually substance dualists.

Descartes claimed that mind and body causally interact with one another. Walking, talking and other bodily movements, are caused by thoughts, decisions and feelings, and we feel pain from physical causes and acquire beliefs from our sense experience. Cartesian substance dualism is most often rejected because it cannot give an adequate account of the causal role of the mind. The objection focuses on explaining the causation of physical events by mental events.

**The conceptual interaction problem**

Nothing seems more obvious than that the mind and the body interact with each other, e.g. I decide to phone a friend and move my body to do so. But how is it that a mental substance, which is not in space and has no physical force, can affect a physical substance, which is in space and moved by physical forces? This puzzle was expressed to Descartes by Elisabeth, Princess of Bohemia. In a letter to Descartes in May 1643, she posed this objection in terms of pushing and movement.

P1. Physical things only move if they are pushed.

P2. Only something that is extended and can touch the thing that is moved can exert such a force.

P3. But the mind has no extension, so it can’t touch the body.

C1. Therefore, the mind cannot move the body.

In fact, as Descartes points out in his reply (letter of 21 May 1643), this isn’t an accurate understanding of how things are moved. For example, we might explain why something falls in terms of its weight. But weight doesn’t ‘push’ the object whose weight it is! Weight is the result of the force of gravity on the mass of an object, and gravity is a force of attraction that operates without needing contact between the two physical objects.

But this is all a matter of details. We can generalise from the force of pushing to force more generally. If the mind is just thought, it has no physical force of any kind. In that case, how could it possibly affect the body? (The mind is not very insubstantial matter; we can understand how something very refined, like a gas, can have causal effects.) And the mind is not in space at all. If causation is thought to involve any kind of spatial relationship between cause and effect, the problem is particularly pressing. Clearly nothing can come into a spatial relationship with a mind which occupies no space.

P1. The movement of a physical object is only initiated by some physical force, exerted at some point in space.

P2. If dualism is true, then the mind is not in space and cannot exert any physical force.

C1. Therefore, if dualism is true, the mind cannot cause any physical object to move.

C2. Therefore, either dualism is false or the mind cannot cause (any part of) the body to move.

In a later letter (1 July 1643), Elisabeth says that she accepts, from her own experience, that the mind does cause the body to move. The problem is that experience gives us no indication of how this happens. She continues, ‘This leads me to think that the soul has properties that we don’t know – which might overturn your doctrine … that the soul is not extended …. Although extension is not necessary to thought, it isn’t inconsistent with it either’. So, we can continue the argument:

P3. The mind can cause the body to move.

C3. Therefore, dualism is false.

But perhaps it is a mistake to try to understand the mind’s power to act on physical objects in terms of how physical objects act on each other. We have a tendency to conceive of all causation in terms of the causation of physical events by other physical events. But perhaps this is mistaken.

Then how should we think about mental causation? Certainly, we can reflect on the fact that we can move our bodies at will. But as Elisabeth points out, the question remains how, according to dualism, this is possible.

The challenge is just as daunting when thinking about how physical objects could cause changes in the mind. How can something which is not thought or consciousness bring about changes in a substance that is entirely thought and consciousness? Physical causation operates, as we said, through the exertion of forces at particular points in space. But it seems impossible to exert a physical force on a mental substance which has no spatial location.

**The empirical interaction problem**

Interactionist substance dualism also faces some empirical challenges. The first is very general. The law of the conservation of energy states that in any closed system, the total amount of energy in that system remains unchanged. The energy can only change forms, e.g. movement can produce heat. A ‘closed system’ is simply one that doesn’t interact with anything outside itself. The universe is usually understood as a closed system, because there is nothing ‘outside’ the universe that it can interact with. So the total amount of energy in the universe can’t change. If something in the universe, such as your body, moved without that energy coming from some other physical source, the law of the conservation of energy would not be true of the universe. So:

P1. If the mind, as a non-physical substance, moved the body, the total amount of energy in the universe would increase.

P2. If the total amount of energy in the universe increased, the law of the conservation of energy would not apply to the universe, and the universe is not a closed system.

C1. Therefore, if the mind moved the body, the law of the conservation of energy would not apply to the universe, and the universe is not a closed system.

C2. Therefore, because the mind, which changes the physical energy in the universe, is not itself physical, physics cannot give us the complete, correct account of physical energy in the universe.

While we may want to say that physics doesn’t tell us everything about what exists, interactionist dualism entails that physics isn’t even the correct account of what exists physically. We can make this more specific to link it to the conceptual issues above: physics is wrong to think that physical movement can only be caused by a physical force.

The second empirical challenge is much more specific. Current science indicates that movements of the body are caused by physical events in the brain. So, if the mind moves the body, it does so by changing what happens in the brain. We may object that we have no evidence of the mind changing what happens in the brain.

That is true, but we have no evidence that the claim is false either. This is because, while neuroscience is making good progress, we still have no clear account of the very complicated causation involved in something like making a choice. But we may think that neuroscience could discover the complete story in time. If interactionist dualism is true, then it seems that what it must discover is that some events in the brain have no physical cause, because they are caused by the mind.

It is common, but perhaps a mistake, to think that there is empirical evidence against substance dualism. The issues are so complex – how does the brain work? Is the universe a closed system? – that we don’t yet have definitive evidence one way or another. So both the objections presented focus instead on the incompatibility between interactionist dualism and empirical science.

**Epiphenomenalist substance dualism**

We could accept that the objections above show that, if substance dualism is true, then mental causation is impossible. But this doesn’t undermine substance dualism if we accept epiphenomenalism, the view that the mind has no causal powers. (An ‘epiphenomenon’ is a by-product, something that is an effect of some process, but with no causal influence.) On this view, the mind does not cause any physical events, so there is no problem of how they do so. But nor do mental events cause other mental events. (For example, the theory of ‘occasionalism’ claimed that nothing has causal powers except God. Whenever it seems like a thought causes me to say something, or a desire causes me to act, what actually happens is that God brings about the effect – my having the thought or the desire is merely the ‘occasion’ for God’s action.)

While it is more common for property dualists than substance dualists to be epiphenomenalists, it is worth briefly considering two objections that can be raised against epiphenomenalist theories at this point.

The first is that epiphenomenalism simply doesn’t accord with our experience. Surely we experience causal connections between our mental states, and between our mental states and our behaviour. It is part of my experience that whether I feel pain makes a difference both to what I think (e.g. that I’m in pain) and to what I do (e.g. jump around shouting). Similarly, when I say what I think, it is part of my experience that my thought is the cause of what I say. It is part of the ‘phenomenology of our mental life’ that we interact with the physical world outside us, that mental processes unfold over time, that our thoughts, feelings, etc. respond to one another and have effects. Epiphenomenalism has to argue that my experience is completely misleading in this respect, because there are no causal connections between my pain, thoughts, and other mental states or events and anything that follows them.

Second, epiphenomenalism makes it hard to understand how we have knowledge of our own mental states. How do I know that I am in pain when I am? The obvious answer is that my belief that I am in pain is caused by my pain itself. I can tell that I am in pain just from introspection. But if epiphenomenalism is true, pain doesn’t cause anything, even my belief that I am in pain. This threatens a natural account of our knowledge of our mental states. If my thoughts and feelings don’t cause my beliefs about my mind, then I could have those beliefs whatever my mental states, just as long as the causes of my beliefs (whatever they are – perhaps brain processes, perhaps God) operate in the same way. In other words, whatever causes me to belief that I am in pain could cause me to have this belief even when I am not in pain. And so my beliefs about my mind, therefore, are unjustified and unreliable. So I can’t know my own mind.

**3. Physicalism**

The view that the mind and the body are separate substances is known as *substance dualism*. Substance dualism claims that there are two fundamental *kinds* of substance – mental and physical. The most common alternative to substance dualism is the view that there is only one kind of substance, which is matter. Thus the mind is not a distinct substance; it is not ‘ontologically distinct’ from what is material, not a separate thing from the body. The claim that there is only one kind of substance, physical substance, is often called ‘materialism’.

In recent years, talk of ‘materialism’ has been supplanted by talk of ‘physicalism’. The most important reason for this is that physics has shown that ‘matter’ is too crude an identification of the most basic substance that exists, e.g. matter can be changed into energy. But in rethinking materialism, philosophers have also refined the claim from being just about what type of substance exists to include other conditions as well. It is not enough that the only *substance* is physical. The fundamental *nature* of the universe is physical, and this covers events and properties as well.

As a first attempt, we could define physicalism as the view that everything that exists – every substance, every property that substances have, every event that occurs – is either physical or completely depends upon something that is physical. ‘Physical’ means something that comes under the laws and investigations of physics, and whose essential properties are identified and described by physics.

But we should be more precise. Physicalism claims that what is physical is metaphysically fundamental. So physicalism says:

1. the properties identified by physics form the fundamental nature of the universe;
2. physical laws govern all objects and events in space-time;
3. every physical event has a physical cause that brings it about in accordance with the laws of physics. (This is known as the ‘completeness of physics’ or ‘causal closure’.)

It is worth saying more about the first and third claims.

The third claim states that all physical events have sufficient physical causes. Of any event involving a change in physical properties (e.g. every movement of your body), that event can be brought about by something physical alone. No other, non-physical causes are necessary. So if there are non-physical causes, they don’t contribute anything *in addition* to physical causes to the way the physical world changes over time.

The first claim states that the properties identified by physics are ontologically ‘basic’. Other properties, in particular mental properties, are ontologically dependent on the properties identified by physics (or more broadly, the natural sciences). Mental properties, therefore, if they exist at all, are not part of the *fundamental* nature of the universe, but ontologically dependent on other properties.

There are three different ways in which this could be true.

Elimination: mental properties don’t exist. The theory eliminative materialism claims that at least some mental properties, as we usually think of them, don’t exist. At least some of our basic concepts of mental properties, such as consciousness or Intentionality, are fundamentally mistaken – these concepts don’t refer to anything that exists. (And any other mental properties that do exist are physical properties.)

Identity: mental properties are, in fact, just types of physical properties. For example, they could be neurological properties. This is the view of mind–brain type identity theory. Neurological properties of the brain, such as what brain cells are made of, the connections they form with each other, the chemicals they exchange, are physical properties. They depend on other more fundamental physical properties to do with molecules and atoms that physics investigates. So mental properties are neurological properties, which are physical properties that depend on more fundamental physical properties.

Dependent but distinct: mental properties are not physical properties of the brain, but they completely depend upon physical properties (perhaps even just physical properties of the brain). But what is it to say that mental properties ‘depend’ upon physical properties? Philosophers spell this out in terms of the idea of ‘supervenience’.

**Supervenience**

The essence of supervenience is this: properties of type A supervene on properties of type B just in case any two things that are exactly alike in their B properties cannot have different A properties.

For example, a painting has various aesthetic properties, such as being elegant or balanced. It also has various physical properties, such as the distribution of paint on the canvas. The aesthetic properties supervene on the physical ones, because we cannot change the painting’s being elegant or balanced without changing the distribution of paint on the canvas. There can be no change in aesthetic properties without a change in physical properties. And two paintings exactly alike in their physical properties (i.e. duplicates) will have the same aesthetic properties. If two paintings are completely identical in terms of how the paint is arranged – if they look exactly the same – then they must also be identical in terms of their aesthetic properties. Of two physically identical paintings, one can’t be graceful while the other is awkward. Any differences in their aesthetic properties entail that there is a difference in their physical properties.

We need to notice the strength of this claim. For physicalism, it is not enough to say that if the paintings are physically identical, then they *are* aesthetically identical. Suppose we say simply that in this case, as it happens, they are both graceful. This allows that in another case, one could be graceful and one not. But that means that aesthetic properties would be able to vary even as the physical properties remained the same. It allows that the physical properties don’t *fix* the aesthetic properties.

This isn’t right. We want to say that if the paintings are physically identical, then they *must be* aesthetically identical. It is not merely false but *impossible* that one is graceful while the other is awkward, if they both look exactly the same. Put another way, once the physical properties of a painting are finalised – when the painting is finished – there is no further work to be done to ‘add’ the aesthetic properties. They are already part of the painting. To change the aesthetic qualities, you *must* change the physical properties.

According to physicalism, physical properties ‘fix’ *all* the other properties in such a way that it is *not possible* for the other properties to change without changing the physical properties. This is what physicalism means by claiming that everything ‘depends on’ what is physical.

We can picture this with the idea of ‘levels’ of existence that correspond to the different sciences. At the bottom is physics, investigating elementary particles and atoms. Molecules involve complex organisations of atoms, and cells involve complex organisations of molecules. Multicellular organisms involves complex organisations of cells, including organs such as the brain. Physicalism understands psychology as relating to the brain and the interactions of living things as just the next level.

If mental properties supervene on physical properties, then they are fixed by physical properties such that it is impossible for two things to have the same physical properties and different mental properties. According to physicalism, just as two paintings with an identical distribution of paint must have the same aesthetic properties, two people with identical properties of their brain, say, must have identical mental properties.

Someone who denies this, who argues that what someone thinks or believes or feels is not completely dependent on their physical properties denies physicalism. They claim that (whether or not it happens) it *is* possible for two beings to have identical physical properties but different mental properties. They may argue, for instance, that physicalism is false because the mind is a separate substance (substance dualism), or because mental properties are somehow independent of physical properties (property dualism).

Physicalism claims that what is physical is metaphysically fundamental. This means either that mental properties don’t exist at all, or that they are identical with certain physical properties, or that they supervene on physical properties.

**3.1. Hempel’s philosophical behaviourism**

Philosophical behaviourism is a family of theories that claim that we *can* analyse mental concepts in terms of concepts that relate to the body, and in particular, the concept of ‘behaviour’. While other theories in philosophy of mind often focus on questions of metaphysics, e.g. whether mental properties ‘exist’ independently of physical properties, philosophical behaviourism focuses on questions of philosophy of language, and what it means to talk about mental properties in the first place. Once we get clear on this, philosophical behaviourism claims, we will see that some of the metaphysical debates about the mind can be avoided. Before we try to do the *metaphysics* of mind, we need to do some *conceptual analysis*.

The term ‘behaviourism’ (without the adjective ‘philosophical’) refers to a theory of how psychology should conduct itself to achieve the status of a science. Science, behaviourism claimed, can only investigate what is publicly accessible. Hence psychology can and must aim only at the explanation and prediction of bodily behaviour, as any talk of or appeal to ‘inner’, inaccessible mental states cannot be scientific. There is no scientific way to establish their existence or nature. This theory, of how psychology should proceed, is *methodological* behaviourism. It makes claims about the methods of science and about *how we can know* about mental states.

By contrast, philosophical behaviourism claims that *what we are talking about* when we are talking about the mind and mental states is behaviour – what people do and how they react. On this view, the mind is not a ‘thing’. Rather, we can talk about organisms ‘having minds’, or better, having mental states, on the basis of how they behave.

There are different kinds of philosophical behaviourism. Although they both agree that we can analyse mental concepts in terms of behaviour, their arguments for philosophical behaviourism, and indeed what they mean by ‘behaviour’, are very different. In this section, we look at the form of the theory defended by Carl Hempel.

Hempel calls his resulting theory ‘logical behaviourism’. Other philosophers sometimes call it ‘analytical’ behaviourism, while the AQA A level syllabus calls it ‘hard’ behaviourism. Confusingly, the term ‘logical behaviourism’ is sometimes also used to mean ‘philosophical behaviourism’, the whole family of theories, and not just Hempel’s original version. We will use the syllabus terms of ‘hard’ and ‘soft’ behaviourism to distinguish Hempel’s and Ryle’s theories.

**The meaning of scientific statements**

Hempel was a member of the ‘Vienna Circle’, the founders of logical positivism who developed and defended the verification principle. In ‘The logical analysis of psychology’, Hempel applied the principle to the question of what psychological language means.

He starts with the general question of what the meaning of a scientific statement is. The answer, he claims, is that to know the meaning of a statement is to know the conditions under which we would call it true and those under which we would call it false. So ‘the meaning of a statement is established by the conditions of its verification’. The ‘conditions of its verification’ are simply the observations that we can make to check its truth. For example, the meaning of the statement ‘the temperature in the room is 21°C’ is given by the (many different) ways in which we can establish whether this is true, e.g. by observing whether ‘the level of mercury in the thermometer in the room is at the mark “21” on the Celsius scale’.

From this account of meaning, we can draw several conclusions. First, if we can’t say what the conditions of verification for a statement are, i.e. if in principle, we cannot empirically check or test the truth of the statement, then it is meaningless.

Second, two statements have the same meaning if they are both true or both false in the same conditions, i.e. if they have the same conditions of verification. If the meaning of the first is given by its conditions of verification, and the meaning of the second is given by its conditions of verification, and the two conditions of verification are the same, then the meaning of the two statements is the same. So ‘the temperature in the room is 21°C’ means the same as ‘the level of mercury in the thermometer in the room is at the mark “21” on the Celsius scale and/or…’ where we fill in the dots by all the other ways we can measure temperature. The statement ‘the temperature in the room is 21°C’ is really just an *abbreviation* of all the statements about its conditions of verification.

Third, this means that we can *translate* a statement into a series of statements that simply describe the conditions of verification. A translation is a statement with the same meaning, but expressed in different words or concepts. We can translate a statement with the concept ‘temperature’ into a series of statements describing the observations we make to establish whether the first statement is true. These statements don’t use the concept ‘temperature’, but concepts of observation and measurement.

**The meaning of psychological statements**

Let’s apply these results to statements in psychology. First, unless we can say how to check whether a statement like ‘Paul has a toothache’ is true or false, it will be meaningless. Second, its meaning is given by its conditions of verification. What might these be? That’s an empirical matter, thinks Hempel, and the list below could be continued, but the conditions of verification will include claims like these:

* ‘Paul weeps and makes gestures of such and such kinds.’ [bodily behaviour]
* ‘At the question “What is the matter?”, Paul utters the words “I have a toothache”.’ [linguistic behaviour]
* ‘Closer examination reveals a decayed tooth with exposed pulp.’ [physical bodily states]
* ‘Paul’s blood pressure, digestive processes, the speed of his reactions, show such and such changes.’ [physiological changes]
* ‘Such and such processes occur in Paul’s central nervous system.’ [brain processes]

What is important about these first two points about the meaning of psychological statements is that psychological statements cannot be about private or inaccessible states of the person. The only way that they could have meaning is if there is some way that we could check whether or not someone has the mental state we say they do. All this means of checking have to be public, so they must relate to physical and behavioural states or changes.

The third implication is that the statement ‘Paul has a toothache’ *means* these claims. It can be translated without loss of meaning into these claims. These are not only ways of checking the truth of the statement, as though such behaviour is a fallible guide to what is privately going on in Paul’s mind. To talk about Paul’s mental states is to talk about Paul’s behaviour and bodily states.

Fourth, these claims describing the conditions of verification don’t use the concept of ‘toothache’ or ‘pain’ or any other mental concept. They only use physical concepts, concepts concerning physical, bodily behaviour and processes. We can generalise the point. All psychological statements can be translated, without changing the meaning of what is said, into statements that only use physical concepts of this kind.

**Implications**

Central to Hempel’s theory is the thought that just as other scientific statements are really abbreviations for statements that describe their conditions of verification, the same is true of psychological statements. This may seem controversial, since we tend to think of psychological and physical concepts as quite distinct. But according to his theory, we can translate statements using psychological concepts into statements using physical concepts. There is no ‘essence’ to mental states and events (e.g. consciousness or Intentionality) that distinguishes them from what is physical. As a result, there is no genuine question about how mind and body relate to one another or interact with one another. Once we correctly understand the logic of mental concepts, such problems disappear.

This isn’t because we have *eliminated* mental states. Hempel’s behaviourism doesn’t say that mental states don’t exist but nor does it say that they do. Instead, the question of their ‘existence’ isn’t a real question. To say that someone is in pain isn’t to say that ‘pain exists’. It is to say that there are certain observations we can make about the behaviour and physical state of the person. The person exists, their body exists, and they behave in certain ways. There is no further question about whether mental states exist.

**3.2. Ryle’s philosophical behaviourism**

**Substance dualism makes a ‘category mistake’**

In *The Concept of Mind*, Ryle argues that the traditional metaphysical framework for discussing the mind rests on a mistake. Cartesian dualism understands the mind in terms of substances and properties. It claims that mind and body are different substances, and that just as there are physical properties and processes, so there are mental properties and processes. Ryle argues that this way of understanding the mind is mistaken.

Ryle calls substance dualism ‘the dogma of the Ghost in the Machine’. The mistake that it makes, he argues, is a ‘category mistake’. What does that mean? Suppose someone is shown around Oxford University – they see the colleges, the buildings with the different faculties and departments, the administrative buildings. But then they ask, ‘I’ve seen the colleges, the faculties, the administration. But where is the university?’ They have misunderstood the concept of ‘university’, thinking that the university is another thing, alongside the colleges, faculties and administration. The person has made a category mistake. The university is not like this; it is how everything that the person has seen is organised.

Or again, suppose someone is having a game of cricket explained to them. The bowler, batters, wicketkeeper and fielders are all pointed out and their tasks explained. But then the person says, ‘I’ve heard a lot about the importance of team spirit. Who does that?’ They have misunderstood the concept and made a category mistake. The exercise of team spirit is not another task like bowling or fielding, nor is someone who is bowling and exercising team spirit doing two separate things. Team spirit is about how the players play the game together.

Concepts belong to different logical categories – different ways in which it makes sense to use a concept. A category mistake is to treat a concept as belonging to a different logical category from the one it actually belongs to.

According to Ryle, substance dualism makes the category mistake of thinking that the mind is like the body – another ‘thing’, a distinct, complex, organised unit subject to distinct relations of cause and effect. The mistake is to think that physical and mental concepts operate in the same way, in the same logical framework of ‘things’ and ‘causes’, ‘substances’ and ‘properties’. But to ‘have’ a mind is not to be in possession of a thing, so that if you have a mind and a body, you have two things. (And ‘losing your mind’ isn’t like losing your keys!) Similarly, talk of mental states and processes understands ‘states’ and ‘processes’ along the lines of physical states and processes. But believing something is not a state in the same sense as the physical state of being solid, and doing mental arithmetic is not a process in the same sense as the physical process of a log burning.

Around the time of Descartes, science reached the stage of plausibly claiming that all physical processes could be explained in non-rational, mechanical terms. So the question arose, what is the place of the mind? Ryle argues that people mistakenly inferred that mental concepts, if they don’t characterise physical processes, must refer to non-physical, non-mechanical processes which occur in non-physical substance. He calls this the ‘para-mechanical hypothesis’. But just as ‘Oxford University’ doesn’t refer to another thing alongside along the buildings and faculties, and ‘team spirit’ don’t refer to another activity alongside bowling, batting and catching, mental concepts aren’t like physical concepts, only applied to a separate thing called ‘the mind’. Instead, we need to think again about the logical analysis of mental concepts – what do they mean, how do we use them?

**Dispositions**

If philosophical behaviourism analysed mental concepts just in terms of actual behaviour, it would be open to an objection. First, we can, to some extent, control our behaviour, e.g. I might stop myself from showing that I am in pain. Second, many mental states, such as knowledge, are dispositions, rather than occurrences. They don’t occur at a time, like actual behaviour does. For example, someone who knows French knows French even when they are talking or reading in English.

So we need to understand ‘behaviour’ not just in terms of actual behaviour, but behaviour that someone would display under different conditions. I want to say that someone *now* understands French (even when *now* they don’t meet Hempel’s conditions of verification), because, e.g. if I *did* ask them whether they speak French, they would answer ‘yes’, or if they *were* in France, they would converse with people there in their own language, and so on. Ryle argues that to talk of mental states and processes is to talk not only of actual behaviour, but also of ‘dispositions’ to behave in certain ways.

Central to Ryle’s argument is his observation that we often speak of mental states in action, in their expression in behaviour. To know how to play chess is something demonstrated in actually playing chess, and we attribute this knowledge to someone on the basis of what they do. Or again, to do something intelligently or thoughtfully – playing, reading, cooking, arguing, etc. – is to be able to regulate what you do. So some of our mental concepts identify skills. A skill isn’t an act – you can’t tell from one piece of behaviour whether it is skilful or just lucky or something else again. But a skill isn’t some invisible, non-spatial thing either (nor, we may add, a physical property of the brain). It is a disposition or complex of dispositions.

What is a disposition? A disposition, in its simplest form, is simply how something will or is likely to behave under certain circumstances. For instance, sugar is soluble. Solubility is the disposition to dissolve when placed in water. Having a disposition is not the same as behaving in a certain way *now*. Sugar is soluble even when it isn’t actually in water. We can express dispositions using ‘if…then…’ statements – hypothetical conditionals. To say that sugar is soluble is to say that *if* sugar is placed in water, *then* it will dissolve.

Solubility is a ‘single-track’ disposition – it is ‘actualised’ or ‘manifest’ in just one way, namely dissolving in water. Other dispositions, such as being hard, have many different ways in which they are actualised. We can infer many different facts from knowing that something is hard; e.g. about whether we can pass other things through it, what sound it will make when hit, whether we can change its shape easily, and so on. We need a series of hypothetical (if…then…) statements to express the disposition of being hard.

Many mental concepts are also concepts of dispositions, so that when we talk of someone having a certain mental state, like being proud or believing that the earth is round, we are talking of what they would do, could do, or are liable to do, in particular situations or under particular conditions, including conditions that they are not in at the moment. Mental concepts can refer to very complex dispositions, dispositions which are ‘indefinitely heterogenous’. For example, in saying that someone is proud, consider the many different and subtle ways in which people can manifest pride (Ryle refers to Jane Austen’s novel *Pride and Prejudice*).

Whether someone has a particular disposition is a matter of whether certain statements about what they could or would do are true or not. These are hypothetical statements, conditional statements of the form ‘if circumstances *c* occur, the person will do *x*’. They are not ‘categorical’ statements that say how things actually are; e.g. many of those circumstances may never arise. They don’t describe actual states of some mental substance. So ‘the mind is not the topic of sets of untestable categorical propositions [as substance dualism must hold], but the topic of sets of testable hypothetical and semi-hypothetical propositions’.

By contrast with Carl Hempel’s ‘hard’ behaviourism, Ryle does not think that statements using a mental concept, such as ‘he is proud’ or ‘he knows French’, can be ‘reduced’ in meaning to a series of hypothetical statements about what the person will do in different situations (or what his physical state is). The mental concept can be analysed in terms of such statements – this is what it means – but we can never give a complete translation, so that we can replace the mental concept by physical ones. Dispositional statements are ‘open’. They support and justify certain inferences, explanations and predictions. To say that someone is proud enables us to draw inferences about how he will behave in certain situations. But we cannot draw all possible inferences and replace the concept ‘proud’ with this set of inferences.

**A note on other minds**

According to philosophical behaviourists, talking about mental states is just talking about actual behaviour and dispositions to behave in certain ways. From how someone behaves, we can infer what behavioural dispositions they have. But from this, we don’t then *infer* that they have a mind. The link between behaviour and minds isn’t based on evidence, it is logical (conceptual). To say someone behaves in certain ways and has certain behavioural dispositions *just is* to say that they have certain mental states. To understand what others say and do is to understand that they have minds. We can know that other people have minds, because we can know directly that they behave in particular ways. Thus, philosophical behaviourism solves the problem of other minds.

**Thinking and other mental processes**

Philosophical behaviourism is on its strongest ground when talking about the mind in action. But what, we may object, about *just* thinking, without acting (which is where Descartes started)?

Ryle’s response to this challenge is first to note that there isn’t just one kind of ‘thinking’. Again, thinking is often done in, with and through action. When we act thoughtfully or intelligently, the thinking isn’t a separate process from the doing, so that the thinking takes place in the mind and the doing in the physical world. There is one process – behaving (reading, driving, conversing …) intelligently – and what makes it an expression of thinking is that it has a certain manner which can be expressed by dispositional statements about what we can, could and would do in certain situations.

But there is also the matter of thinking quietly ‘to oneself’. Ryle’s central claim here is that this is *internalised speaking*: ‘Much of our ordinary thinking is conducted in internal monologue or silent soliloquy’. Speaking is, of course, an overt behaviour, and we only acquire the ability to think – to speak silently to ourselves – with effort. The silence, and the fact that we are speaking only with ourselves, are *inessential* to the nature of thinking. To think through a maths problem, one can do so either with pen and paper, articulating the steps as one goes, or silently, ‘in one’s head’. Whether a process is public or private is irrelevant to whether it is thinking. ‘The phrase “in the mind” can and should always be dispensed with’. Mental processes only sometimes and only contingently take place ‘in the mind’. Processes that do, as it happens, take place silently don’t define thinking any more than those that take place as publicly observable behaviours.

**Dispositions and occurrences**

Thinking is something that happens at a time and takes time. It is a process, it ‘occurs’, it is a mental ‘occurrence’. So we can’t say that thinking is *just* a matter of dispositions. The same is true of other mental occurrences and processes, such as being conscious of (paying attention to) what you are doing, feeling or thinking (what Ryle calls ‘heeding’). What’s the relation between occurrences and dispositions?

To understand this, compare ‘it is dissolving’. This states that something is happening, but does so in dispositional terms. From ‘it is dissolving’, we know that it is soluble, and so dissolves in water. So it would do just what it is doing in this situation, given that it has that disposition.

Likewise, to say that someone is paying attention to what they are doing is to attribute dispositions about what they could say if you asked them, but also to add that they are ‘in the mood or frame of mind’ to do just what it is that they are doing. This is what Ryle means by a ‘semi-hypothetical’ statement – it both explains an actual occurrence and enables us to make inferences.

**A note on physicalism and the category mistake**

In the section ‘Substance dualism makes a ‘category mistake’’, we saw that Ryle argues that dualism misunderstands the logic of mental concepts. It understands the mind as another ‘thing’, like the body in fitting into a metaphysical framework of substances, properties, and causation. We can extend his criticism to some physicalist theories as well.

Some physicalist theories, such as type identity theory and eliminative materialism, reject the idea that the mind is a separate substance, they understand mental properties and physical properties in the same way. Mental properties *are* physical properties, according to type identity theory. According to eliminative materialism, mental properties are part of an empirical theory that offers causal explanations, just like other scientific theories of the physical world. However, unlike dualism, which infers that mental concepts refer to non-physical, non-mechanical processes, these physicalist theories infer that mental concepts must refer to the same physical, mechanical processes that our physical concepts refer to.

Ryle’s philosophical behaviourism rejects both options. An analysis of our mental concepts shows that they don’t work like physical concepts. While physical explanations use categorical concepts, mental concepts are dispositional.

So is philosophical behaviourism a form of physicalism? We can say that it is, because, according to philosophical behaviourism, there is no distinct psychological ‘reality’ – no distinct psychological substances or properties. This isn’t because the theory eliminates them, but because questions about the mind aren’t questions about what exists. What exists is given by natural science. Categorical facts about substances, their properties and causes belong here, in the descriptions of the world that natural science provides. Dispositions depend on such categorical facts – sugar’s disposition to dissolve depends on its physical properties, and our dispositions to behave as we do depend on our physical properties. But dispositions, for Ryle at least, aren’t additional ‘properties’ (at least, of the same kind as physical properties). Dispositions are expressed in hypothetical statements, not categorical ones. And saying ‘if this happens, then this will happen’ doesn’t state anything about what exists.

**3.3. Philosophical behaviourism and consciousness**

**The asymmetry between self-knowledge and knowledge of other people’s mental states**

In *The Concept of Mind*, Gilbert Ryle observes that it is part of the ‘official doctrine’ of substance dualism that the ways in which we gain knowledge of our own and others’ mental states are very different. We are directly aware of our own mental states, but we can only infer those of others. Our self-knowledge comes from our consciousness of our mental states and our introspection of that consciousness. We cannot be conscious of anyone else’s mental states in the same way. Furthermore, we are aware of our mental states in such a way that we cannot make mistakes, but this is not true of our beliefs about other people’s minds.

Now if mental states were dispositions to behaviour, or again if what it means to say that someone has a particular mental state is given by conditions of verification, all this wouldn’t be true. Given that what I am saying when I say ‘I am in pain’, or ‘I believe that Paris is the capital of France’, is that I behave or am disposed to behave in certain ways, then it seems that I would have to infer what mental states I have from how I behave, or how I think I am disposed to behave.

But, we can object, this isn’t right. I can know what I believe, what I want or fear or hope, directly, without inference and without thinking about how other people would verify whether I have these mental states. Furthermore, if I am thinking to myself, I know what I am thinking in a way that no one else can.

P1. The analysis of mental states in terms of behavioural dispositions (or conditions of verification) rules out an asymmetry between self-knowledge and knowledge of other people’s mental states.

P2. Yet it seems obvious from experience that there is such an asymmetry.

C1. Therefore, philosophical behaviourism is false.

**Ryle on consciousness**

Self-knowledge is not an issue that Carl Hempel addresses in his explanation of behaviourism in ‘The logical analysis of psychology’. However, one response he could make is that he is only interested in discussing the meaning of statements in psychology, which as a science, deals only with knowledge of other people’s mental states. But this response is unsatisfactory, since we obviously do use mental concepts when talking about our own mental states. The challenge remains how Hempel’s behaviourism can understand and explain self-knowledge.

Ryle’s response is to argue that consciousness, understood in as giving special self-knowledge through introspection, is a myth. He argues that self-knowledge and our knowledge of other minds is on a par, gained in the same way in each case – by paying attention. This enables us to make reliable dispositional claims about our own or other people’s behaviour, whether this is overt or silent. The main difference is simply that we have more evidence available to ourselves.

Central to Ryle’s argument is that being conscious of something is to pay attention to it (to ‘heed’ it, he says). We can pay attention to what we are doing and to what we have just felt or said silently to ourselves. But we can also pay attention to what someone else is doing and what they say out loud to us. To know what we are thinking or feeling is not to stand in some special, inner private relation to certain mental ‘objects’ (‘thoughts’, ‘feelings’), but for us to be ready to say what we think or feel and be unsurprised by the occurrence of the thought or feeling.

Compare not knowing a process in one’s mind: you make a joke spontaneously or come up with a solution to a problem. How did you do it? You can’t say – the joke or solution comes as a ‘surprise’ to you. Knowing what you are thinking or doing is just to be continuously prepared for what comes next in that process.

Hence consciousness provides the same kind of knowledge in cases of knowing our own mental states and knowing the mental states of others. The main difference is that in our own case, we have more to go on, because we are the audience of our silent, inner speech – our thinking – and others are not.

Knowing what you are thinking is not different in kind from knowing what someone else is thinking, since we can know just as directly what someone else thinks when they speak, at least when they speak in an unguarded, unembarrassed and uncalculated way, which is the most natural way to speak. When we talk like this – whether to others, or silently to ourselves – we are directly expressing our mental states. So when we pay attention to what we say, we gain knowledge of the mind of whoever is talking. Introspection is not a form of perception of special mental objects. It is just to pay this kind of attention to ourselves.

**Objections to Ryle’s theory of consciousness**

We can make two objections to Ryle’s analysis of consciousness and self-knowledge. First, can thinking be adequately understood in terms of inner speech, and can internalised speech form the model for mental processes generally? What about non-linguistic mental processes or changes in feeling and mood? Here we aren’t saying anything to ourselves.

Second, Ryle seems to miss out the subjective, experiential aspect of mental states and processes. The distinctive quality of certain experiences, e.g. how a sensation or emotion feels to the person experiencing it, is central to our mental lives. I have an awareness of this aspect of mental states and processes for my own mental states, but not for anyone else’s. How does philosophical behaviourism account for this aspect of self-knowledge and the asymmetry between self-knowledge and knowledge of other people’s mental states?

It may be that these objections don’t themselves re-establish a strong asymmetry between self-knowledge and knowledge of others’ mental states. But they attack Ryle’s rejection of it.

**The distinctness of mental states from behaviour**

It seems that many mental states and processes have an ‘inner’ aspect that can’t be captured by behaviour and behavioural dispositions. We should agree that to be in pain often involves doing certain things, such as wincing, recoiling from the cause of pain, nursing the damaged part of the body, etc. (call all this ‘pain behaviour’). But this doesn’t capture the ‘essence’ of pain, which is that it hurts.

More generally, we can argue that statements about behaviour or conditions of verification only tell us about how to know about mental states. They don’t give us the ‘intrinsic nature’ of mental states, what they are ‘in themselves’, so to speak. Mental states are not the same as these conditions of verification or behavioural dispositions, but are something distinct or something more than them. (For instance, perhaps they cause this behaviour, and that is why they are correlated with it.)

In his discussion of this objection, Hempel responds that our understanding of people in terms of their mental states, e.g. that they are in pain or that they want chocolate, is tied up with their physical state and their behaviour. We can’t understand what it is for them to be in such-and-such a mental state without referring to such physical conditions. What would it be for someone to want chocolate but never seek it or express this in language? What is a ‘desire’ if not something that motivates behaviour? The behaviour isn’t just evidence of their mental state, it gives us the meaning of the concept. Furthermore, our understanding of them is based on the information we have about them – and all the information we get concerns their bodily behaviour. Because meaning is given by conditions of verification, we can only meaningfully talk about things that we can gain information about, and the meaning of a concept is given by the ways in which we can check the truth of claims that use it.

But what, we may object, about perfect actors? Someone can pretend to be in pain, and may do so utterly convincingly, and yet not be in pain. Or again, someone may ‘live a lie’, pretending to have certain beliefs and desires without actually doing so.

Ryle might respond that mental states aren’t just doing certain things, but to have the disposition to do them. The actor doesn’t have the same dispositions that someone who really feels pain has. There are ‘if…then…’ statements that are true of the actor that are not true of the person who is really in pain. And similarly with someone pretending to believe or want what they do not, in fact, believe or want.

This response may be convincing for Intentional mental states such as beliefs and desires. But we can object that, at least when it comes to phenomenal properties of consciousness, this analysis misses an important point. Pain isn’t just a disposition to shout or wince; there is also how pain feels, ‘what it is like’ to experience pain. This is what distinguishes the person in pain from the actor. It is highly counter-intuitive to argue that this aspect of experience is constituted entirely by behavioural dispositions.

To make the point, in his article ‘Brains and behaviour’, Hilary Putnam asks us to imagine a community of ‘super-spartans’. (The Spartans were an ancient Greek community who were very tough and discouraged demonstrations of pain.) These are people (or creatures) who so completely disapprove of showing pain that all pain behaviour has been suppressed. They aren’t acting; this is how they are in everyday life. They no longer have any disposition to demonstrate pain in their behaviour. Yet, they could still be in pain. Pain is conceivable without any associated pain behaviour. Pain can’t be understood just in terms of dispositions to pain behaviour, it is distinct from such behavioural dispositions. So philosophical behaviourism is false.

Hempel argues in response that actors meet only some of the conditions of verification for pain, those based on directly observing their behaviour in the ordinary sense of the word. But his behaviourism claims that ‘behaviour’ covers physiological and neurological states and processes as well. And here there will be a difference between actors and people who are genuinely in pain. The physiology and brain activities of actors will be different. He gives the example of mental illness. No one can have all the symptoms of being mentally ill and yet not be mentally ill. If they could, then there would be no difference between being mentally ill and not being mentally ill. There must be something that distinguishes people who are mentally ill from those who are not (including actors).

We can apply his response to Putnam’s example of super-spartans. There will be conditions of verification for saying that a super-spartan is in pain. Given that they do not show pain in their overt behaviour, the conditions of verification will have to prioritise statements about their physiology and brain processes (or whatever physical processes underlie their pain response).

Such a response would accept that mental states are distinct from behaviour in the ordinary sense of the word, but not distinct from the physical states of the body. However, if Hempel prioritises the physiological and neurological conditions of verification over overt behaviour, then his theory starts to sound more like a kind of type identity theory (though one that proposes an analytic reduction of mental states to physical states rather than an ontological one).

**3.4. Objections to Behaviourism**

**Is mind without body conceivable?**

If philosophical behaviourism provides the correct analysis of mental concepts, then it is inconceivable for there to be a mind without a body. A mind is not a thing, it does not ‘exist’ in the same way as bodies exist only with different properties. To think of the mind as a thing (and hence something that could exist in its own right) is a category mistake, and category mistakes are misconceptions. Whether we use Ryle’s ordinary concept of behaviour or Hempel’s physical one, to behave requires a body. As dispositions to behave, mental states can only be had by creatures that can behave in certain ways. Or again, as statements about behaviour, psychological statements can only be true of creatures that can behave in certain ways.

Given this, if we can succeed in showing that it is conceivable for the mind to exist without the body, then it seems that logical behaviourism must be false. Now, many people have thought that mind without body is conceivable – belief in God and the existence of one’s soul in the afterlife demonstrate this. Shouldn’t the analysis of our mental concepts make such common thoughts coherent rather than incoherent? Aren’t our concepts defined by how we use them?

In *The Concept of Mind*, Ryle argues that such beliefs don’t actually reflect how we use our concepts in everyday life. The ‘official doctrine’ conflicts ‘with the whole body of what we know about minds when we are not speculating about them’. The belief in mind without body is not part of everyday use, but the result of theological and philosophical theorising. We cannot have a clear and distinct idea of ourselves as ‘minds’ only, and we are mistaken if we think that we can. ‘Many people can talk sense with concepts but cannot talk sense about them’.

Hempel would agree. For our mental concepts to have genuine meaning, we must be able to provide the conditions of verification for statements that use such concepts. These conditions of verification must refer to behaviour, since the only information we can have about other people’s mental states is from their behaviour.

**Issues defining mental states satisfactorily**

Philosophical behaviourism claims that we can understand our concepts of what mental states are in terms of behaviour and dispositions to behave. But to provide such an understanding, we need to successfully identify the behaviour that provides the conditions of verification for saying that someone has a particular mental state (Hempel), or say just what behaviour the mental state disposes us toward (Ryle). Or more precisely, even if we don’t need to be able to actually do this, we need to think that, in principle, such an analysis is possible. But is it? There are two reasons to think that it is not.

**The multiple realisability of mental states in behaviour**

The first is the ‘multiple realisability’ of mental states in behaviour. A mental state might be expressed in quite different behaviours not only in different situations, but even in very similar situations by different people. In fear, faced with a lion, I might freeze, you might run. But then again, in fear, faced with a snake, I might run, you might freeze. Or again, on one occasion, I run, and on another, I don’t. ‘Multiple realisability’, in this context, just means that there are many ways in which the disposition (the mental state) can be actualised (expressed in behaviour).

How can we possibly give a list of the conditions of verification for all the ways in which people might behave when afraid? The list will be infinitely long, especially if we have to specify all the different conditions under which people show fear. There is no way we can complete the sentence ‘A person is afraid = if they are in situation x, they will do action A, or if they are in situation y, they will do action B, or if they are in situation z, they will do action C, or…’

The objection from multiple realisability can be understood in two ways. First, it shows that the analysis of mental states in terms of behaviour is not possible. There is no finite set of statements about behaviour which provides an account of the meaning of a mental concept. So philosophical behaviourism does not give an adequate account of what our mental concepts mean.

P1. People with the same mental state behave differently, both in different circumstances and even in the same circumstance.

P2. It is not possible to draw up a finite list of hypothetical conditionals or statements of the conditions of verification that describe all the ways someone with that mental state may behave.

C1. Therefore, the claim that mental states can be analysed in terms of behaviour is false.

C2. Therefore, philosophical behaviourism is false.

Second, if different people with the same mental state have dispositions to do different things in similar situations, how can we say that these different dispositions are actually the same mental state? What is it that makes it the same mental state, given that the dispositions are different? The objection shows that what makes any mental state the mental state that it is – what makes a pain pain, what makes the fear of snakes a fear of snakes, what makes the belief that Paris is the capital of France the belief that Paris is the capital of France – cannot be simply how someone behaves or is disposed to behave. The conditions of verification and/or behavioural dispositions don’t express the identity conditions for mental states.

P1. People with the same mental state behave differently, both in different circumstances and even in the same circumstance.

C1. Therefore, what makes it true that two people have the same mental state is not that they have the same behavioural dispositions.

C2. Therefore, philosophical behaviourism is false.

**Circularity**

The second difficulty in analysing mental states in terms of behaviour and behavioural dispositions is that how someone behaves in a particular situation depends not on just one mental state, such as being afraid, but on how this interacts with other mental states. Suppose I am afraid of dangerous snakes. Does this dispose me to run when I see one? That depends. Do I believe the snake is dangerous? Do I believe that this type of dangerous snake is one you shouldn’t run away from? Am I able to recognise the type of snake? Do I want to avoid being bitten? And so on.

We can’t specify the conditions of verification for fear, or again, what set of dispositions fear is, without mentioning other mental states. That’s a problem for the claim that we can analyse mental states in terms of behaviour. Suppose I want to define the concept ‘furniture’. I say ‘furniture is tables, chairs, bookcases, and other pieces of furniture’. This is an awful definition, because the term ‘furniture’ appears in my definition of what ‘furniture’ means. The definition is circular, because the term we want to define appears in the definition.

The same challenge faces philosophical behaviourism. If we try to say what fear of snakes is by including sentences like ‘if someone recognises a snake and believes that the snake is dangerous, then they will run’, while this doesn’t mention ‘fear’, it mentions other mental states. If we then provide a further analysis of these mental states, such as the belief that the snake is dangerous in terms of dispositions as well, we will have to mention other mental states again. What behaviour my belief that the snake is dangerous disposes me towards will depend on other mental states. In fact, it will depend on whether or not I am also afraid of snakes!

The objection is that there is no way of analysing a mental state in terms of behaviour (either in terms of conditions of verification or in terms of dispositions) without mentioning other mental states. And so the analysis will be circular. A circular analysis, like a circular definition, is unsatisfactory. Philosophical behaviourism aims to tell us what mental states are, but can’t do so without talking about mental states in the analysis! So it doesn’t provide an analysis of what mental states are after all.

**Hempel’s response to multiple realisability and circularity**

The objections from multiple realisability and circularity are particularly forceful against Hempel’s hard behaviourism, since he claims that the conditions of verification give a complete translation of statements using mental concepts in terms statements about behaviour. For the translation to be complete, then we need a finite list of the conditions of verification. But multiple realisability suggests that there can be no such list. And for it to be a translation, then we must avoid using mental concepts in stating the conditions of verification. But circularity suggests that we cannot eliminate mental concepts in this way.

One response Hempel could make is to emphasise the importance of statements about physiology and brain processes. While people may behave in many different ways in different situations, their physiology and brain processes will be the same, he could argue. And it is these, not the many varied statements about how people might act, which are central to identifying what mental concepts really mean.

However, this response would move his theory closer to a form of type identity theory, since ‘behaviour’ turns out not to be as important as physical properties of the body. As a form of type identity theory, it doesn’t avoid the problem of multiple realisability, since type identity theory faces its own version of this problem! The physiology and brain processes of different people (or certainly, of different species) could well be very different, even though they have the same mental state.

In fact, Hempel eventually abandoned his theory. He gave up the claim that we can define mental concepts in terms of behavioural conditions of verification; indeed, he gave up verificationism. Instead, he accepted that scientific statements as a whole, including psychological ones, may introduce ‘hypothetical entities’, e.g. beliefs, genes, atoms, and so on. Claims about such entities cannot be understood just in terms of how we verify them; the relationship between talk of such things and testing the truth of claims about them is more complicated than that.

**Ryle’s response to multiple realisability and circularity**

While Hempel’s theory must answer the objections, Ryle’s theory may escape them, because they misunderstand what he is claiming. In essence, he accepts both points, and builds them into his theory.

First, in accordance with circularity, Ryle argues that disposition statements are ‘open’, and cannot be replaced by a complete set of hypothetical statements linking particular matters of fact (such as a situation and a behaviour). Therefore, Ryle accepts that it is impossible to specify mental states in terms of dispositions, replacing mental concepts with behavioural ones alone in our thought and language. Second, in accordance with multiple realisability, Ryle argues that mental concepts are concepts of ‘indefinitely heterogenous’ sets of dispositions. Nothing that is ‘indefinite’ can be exhaustively characterised – no finite list is possible. The objections are correct, but they are not objections, since his philosophical behaviourism doesn’t aim to offer finite translations of psychological claims.

However, this response doesn’t address the second interpretation of the objection from multiple realisability. If mental states are so ‘heterogenous’, what makes a mental state the mental state that it is? Is there any stable correlation (let alone identity) between a mental state and behavioural dispositions? It seems that, given circularity, a particular mental state could be (compatible with having) a disposition to just about any behaviour, depending on a person’s other mental states at the time. My fear of dangerous snakes could dispose me to say ‘Well, hello there, Mr Muggins!’ if I also believe that this phrase effectively prevents snake attacks!

If mental states are behavioural dispositions, then mental state x = behavioural disposition x and mental state y = behavioural disposition y. A different behavioural disposition is a different mental state, and the same mental state must give us the same behavioural disposition. But according to multiple realisability, you and I could have mental state x, but I have behavioural disposition x while you have behavioural disposition y, given our other mental states – we are both afraid of the snake, but I am disposed to say ‘hello there, Mr Muggins’ and you are disposed to run! Leibniz’s law of the indiscernibility of identicals says that if two things are identical (i.e. are just one thing), then they must share all the same properties. So if we have the same mental state but different behavioural dispositions, by Leibniz’s law, mental states can’t be behavioural dispositions.

Ryle can reply that we shouldn’t focus on individual ‘pieces’ of behaviour that may differ, as this again misunderstands the theory. Ryle’s analysis allows that we can’t tell what disposition, if any, is being expressed in a single piece of behaviour. What makes the behaviour the expression of the disposition that it does, in fact, express, depends on whether certain hypothetical statements about other situations are true or not. So whether or not I feel afraid of the snake isn’t fixed only by what I actually say or do, but by a whole host of dispositions to other behaviour in other circumstances, e.g. what I would do if the snake comes swiftly towards me after I said ‘hello there, Mr Muggins’, and so on.

Once we recognise this, we can say that while mental states involve ‘indefinitely heterogenous’ dispositions, we shouldn’t overstate the case. On the whole, people in the same mental state have very similar dispositions. Many of these similarities hold even when there are some things they do differently. For example, for fear, there are similarities in how they answer ‘are you scared?’, their facial expressions, etc. There is sufficient overlap for us to say that they are in the same mental state.

Furthermore, Leibniz’s law is about identity in metaphysics, and so it doesn’t obviously apply to Ryle’s theory. His claim that mental states are behavioural dispositions isn’t a claim of identity between two ‘entities’, but a logical analysis of what we mean when we are talking about mental states. Despite not being able to reduce mental concepts to statements about behaviour, Ryle argues that a concept of a mental state is a concept of a set of behavioural dispositions. Ryle’s philosophical behaviourism provides an analysis of the meaning of mental concepts, but it does not justify the claim that we could replace talk of mental concepts with talk of specific behavioural dispositions. Mental concepts work at a higher level of generality that can’t be reduced to sets of individual hypothetical statements about behaviour.

**3.5. Mind – Brain Type identity theory**

‘Type identity theory’ claims that mental properties *just are* physical properties. If we say that these physical properties are properties of a brain, then the theory is the mind–brain type identity theory. So this theory claims that thinking a thought or feeling an itch is *exactly the same thing* as certain neurons firing, say, and having a belief is the same thing as, say, certain neural connections existing. Any particular type of mental state is a particular type of brain state.

If we understand ‘physical property’ to cover all the properties investigated by the natural sciences (physics, chemistry, biology, geology, etc), then there are lots of different kinds of physical property. For instance, a swan is a bird and (usually) white – but what makes it a bird (a biological property) and what makes it white (a colour property) are different properties, though both are physical properties in the broad sense intended here. Just as being a swan and being white are physical properties, mental properties are also physical properties, probably highly complex neurophysiological properties, claims mind-brain type identity theory. Just as brains have physical properties of size and weight, their neurophysiological properties – synaptic connections, neurochemistry, the structure and organisation of neurons – are also physical properties. And some of these neurophysiological properties are thoughts, or pains, or desires… The mind is certain patterns of brain processing interacting with other patterns of brain processing.

Mind–brain type identity theory was developed in the 1960s as neuroscience gathered pace. The evidence is that mental events and states are very closely dependent on the brain, so many people now think that ‘the mind’ is just ‘the brain’, and everything mental is actually neurophysiological. The theory is called ‘type’ identity, because it claims that mental ‘types’ of thing (mental properties, states and events) are physical ‘types’ of thing (physical properties, states and events). Mental ‘things’ turn out to be physical ‘things’; i.e. mental properties are actually physical properties of the brain, mental states are brain states.

Mental and neurophysiological properties may not *seem* the same, but, argues the type identity theorist, that’s because we have different ways of knowing about these properties – through experience and through neuroscience. As we investigate the world through science, many things turn out to be something they don’t seem to be; e.g. solid objects are mostly empty space, water is just hydrogen and oxygen (who’d have guessed?).

**Correlation, identity and reduction**

Type identity theory needs to be distinguished from the claim that mental states are *correlated* with brain states. For example, having a heart is correlated with having kidneys – every animal that has a heart has kidneys and vice versa. But hearts and kidneys are not the same thing! Or again, having a size and having a shape are correlated – everything that has a size has a shape and vice versa. But size and shape are distinct properties. So simply pointing out that everything that has a particular brain state also has a particular mental state doesn’t show that mental states and brain states are the same thing. They could be two distinct things that occur together. Correlation is not identity.

Suppose that neuroscience shows that thoughts, feelings, etc. are correlated with what happens in the brain. Why go on to argue that they are identical? J. J. C. Smart replies, in ‘Sensations and brain processes’, because that is simpler.

Smart claims that ‘[s]ensations are nothing over and above brain processes’ – not correlated, but identical. While he defends the mind–brain type identity theory for sensations, the theory can be generalised to other mental states and occurrences. The motivation for the theory, Smart says, is simply Ockham’s razor. If there are no overwhelming arguments in favour of dualism, then we should reject the idea of distinct non-physical substances or properties. Science indicates that the neurophysiological properties of the brain are a good candidate for what mental properties are.

To understand the theory, it is important to get clear on the identity claim. It is not a claim about language or concepts, but about reality. So the claim is not, for example, that the concept PAIN *means* NOCICEPTOR FIRING (nociceptors are a type of neuron involved in pain). The identity theorist is not offering definitions of the terms that we use. The identity claim is, therefore, not meant to be *analytically* true. ‘Pain is the firing of nociceptors’ (if it is true) is not true in the same way that ‘bachelors are unmarried men’ is true. The concepts PAIN and NOCICEPTOR FIRING remain distinct. The claim is that both concepts refer to *the same thing* in the world. The firing of nociceptors is what pain is. Two concepts, one property.

Understanding the nature of the identity claim helps deal with objections. In particular, we might object that we can talk about sensations without knowing anything about brains. So they can’t be the same thing. But, replies Smart, we can talk about lightning without knowing anything about electrical discharge – because the concepts are distinct. That doesn’t show that lightning isn’t electrical discharge. Or again, water is H2O. The concepts are distinct, but the properties are the same. The same is true of ‘sensation’ and ‘brain process’.

Type identity theory is a type of ‘reductive’ physicalism. An ‘ontological reduction’ involves the claim that the things in one domain (e.g. mental things) are identical with some of the things in another domain (e.g. physical things). For example, we can argue that heat is just mean molecular kinetic energy. They are the same thing. Or again, although they seem different, electricity and magnetism are the same force, electromagnetism. Every mental property, type identity theory argues, is a certain physical property. The identity claim is a reduction because we have ‘reduced’ mental properties – which we might have thought were a different kind of thing – to physical properties. I.e. there is *nothing more* to mental properties than being a certain kind of physical property.

**Obj: Putnam and the multiple realisability of mental states**

The most famous objection to the type identity theory was developed by Hilary Putnam in ‘The nature of mental states’. He argues that mental properties are not *identical* to physical properties because the *same* mental property can be related to or supervene on *different* physical properties. For example, the brain states that relate to pain may well be different in different species, in humans and birds, say, but pain is the same mental state. If this is true, there are creatures who, when they are in pain, have different physical properties from us when we are in pain. Therefore, ‘being in pain’ cannot be exactly the same thing as having a particular physical property. This is the argument from ‘multiple realisability’.

The term ‘realise’ here means ‘to give actual form to’ or ‘to bring into reality’. For instance, we can talk of a design for a dress being beautifully realised in the final product. Similarly, in metaphysics, philosophers talk of one property ‘realising’ another. To say that a particular neurophysiological property ‘realises’ pain in human beings is to say what pain is, the form of existence it has, is given by that neurophysiological property.

Putnam’s argument is that many different neurophysiological properties could realise pain, and so pain can’t be the same thing as any one of those properties. Why not? Well, suppose pain is identical to some neurophysiological property in humans, call it N1. Pain = N1. If they are one and the same thing, then whatever has N1 is in pain, and whatever is in pain has N1. Now suppose pain is identical to some neurophysiological property in dogs, call it N2. Pain = N2. If they are one and the same thing, then whatever has N2 is in pain, and whatever is in pain has N2. If pain = N1 and pain = N2, then N1 = N2 (pain = pain). N1 and N2 must be the same neurophysiological property. But what if dog’s brains process pain in a different way from human brains, so that N2 is a different neurological property from N1? If N2 ≠ N1, and pain = N1, then the dog is not in pain. But the dog is in pain! So if both the human being and the dog are in pain, but they have different neurological properties, then pain can’t *be* exactly the same thing as either N1 or N2. (This still allows that pain can be *correlated* with or *realised* by N1 in human beings and N2 in dogs.)

As Putnam presents it, this is an empirical argument, but it is a very plausible one. It becomes yet more plausible when we consider other mental states and non-terrestrial species. If there are aliens, given that they evolved completely separately from us, if they have mental states, it is extremely unlikely that they will have the same physical states as us. But according to type identity theory, to have a particular mental state just is to have a particular physical state. So the theory is making a very implausible prediction.

The argument can also be rephrased as an a priori argument from conceivability:

P1. It is conceivable, and therefore possible, for a being with quite a different physical constitution from us to have the same thoughts or sensations.

P2. But it is inconceivable, and therefore impossible, for something both to have and not have a certain property.

C1. Therefore, mental properties can’t be the same as physical properties.

The identity theorist could respond that we should talk about ‘human pain’; that this is a different property from ‘dog pain’. Or again, if there are intelligent aliens who have thoughts, but different brains, we should talk of ‘human thoughts’ and ‘alien thoughts’. But this doesn’t seem plausible – pain is pain because of *how it feels*; thought is thought because of *what is thought*. A dog and a human being in pain share something in common, which we identify as the mental property ‘being in pain’. If an alien believes that snow is white, and so do I, we have the same type of thought, whatever our physiology.

(Again, this is not to say that there is *no* relation between mental and physical properties. It is just to argue that the relation is not identity. For instance, we can accept that mental states are correlated with brain states in human beings, while also allowing that in different species, the same type of mental state is correlated with a different type of physical state.)

**3.6. Eliminative Materialism**

Eliminative materialism (also known as eliminativism) argues that future scientific developments, especially in neuroscience, will show that the way we think and talk about the mind is fundamentally flawed, at least in some very important respects. At least some of our mental concepts are so mistaken that they refer to things that neuroscience will show don’t exist. Central to our normal understanding of the mind are phenomenal properties and Intentionality. Eliminative materialism argues that neuroscience will revolutionise our understanding of each so that we may question whether they exist at all as we think of them now. As neuroscience proceeds, it will show that at least some of our central psychological concepts don’t refer to anything – nothing exists that corresponds to some mental terms, e.g. ‘belief’, ‘desire’ or ‘pain’ or even ‘Intentionality’ and ‘consciousness’.

**Patricia Churchland on reduction and elimination**

Reductive explanation

In the simplest terms, an ‘ontological reduction’ involves the claim that the things in one domain (e.g. mental things) are identical with some of the things in another domain (e.g. physical things). For example, we can argue that heat is just mean molecular kinetic energy. They are the same thing. Or again, although they seem different, electricity and magnetism are the same force, electromagnetism. A similar claim could be made concerning mental properties: ‘Every mental property is a certain physical property’. The identity claim is a reduction because we have ‘reduced’ mental properties – which we might have thought were a different kind of thing – to physical properties. I.e. there is *nothing more* to mental properties than being a certain kind of physical property. (This claim is discussed in the handout ‘Mind-brain type identity theory’.) One reason to make such a claim is Ockham’s razor. If mental properties are physical properties, then fewer things exist – the metaphysics is simpler.

But how does science ever come to make such a claim? In *Brainwise*, Patricia Churchland argues that what is involved is not just a simpler metaphysics (the claim that fewer things exist) but a more powerful explanatory theory. Inference to the best explanation goes beyond Ockham’s razor. If identifying two properties as the same thing enables you to explain something that you can’t otherwise explain, that is the best reason for thinking they are the same thing. Otherwise, we can’t move beyond the claim that they are merely correlated.

Ontological reduction is part of reductive causal explanation:

*“a reduction has been achieved when the causal powers of the macrophenomenon are explained as a function of the physical structure and causal powers of the microphenomenon. That is, the macro-properties are discovered to be the entirely natural outcome of the nature of the elements at the microlevel, together with their dynamics and interactions.”*

For example, we can explain everything about water – why it is liquid at certain temperatures, why it is transparent, why cars skid on it, why we can’t breathe in it but fish can, etc. – in terms of the nature of molecules of H2O, how they are structured and how they interact with each other and other things (such as car tyres or our lungs). To ‘reduce’ water to H2O is just to be able to explain all the causal powers of water – the effects it has on other things and the effects other things have on it – in terms of the causal powers of H2O molecules.

The identity claim doesn’t mean that the *concepts* of the macro-theory mean the same as those referring to the micro-properties. WATER doesn’t mean H2O, and THOUGHT doesn’t mean ‘neurophysiological firing pattern x’. However, when one theory offers a reductive explanation of things in another theory, it often happens that the meanings of the concepts *change* in light of new empirical discoveries. For example, the term ATOM meant ‘indivisible fundamental particle’, but then physicists became able to split the atom. So the meaning of ‘atom’ changed to mean ‘the smallest existing part of an element consisting of a dense nucleus of protons and neutrons surrounded by moving electrons’.

Beyond reduction to elimination: the example of heat

However, sometimes the empirical discoveries indicate that rather than changing the meaning of the concept, we should give up on that concept and what it refers to completely. In other words, the concept should be eliminated because nothing exists in the way it supposes.

A good example is given by the history of the science of heat. What do you think heat is, just from everyday experience? Well, hot things have more of it than cold things. Heat passes from hot things to cold things. Hot things ‘give off’ heat. So how about this suggestion: heat is a kind of fluid that makes things hot and can be passed from one thing to another. This was the theory of heat in the late 18th century, and the fluid was called ‘caloric’ (as in ‘calories’).

Ok, so if hot things have more caloric fluid than cold things, they should weigh more. So, when you heat something up, it should get heavier. Scientists tested this. Heating something up doesn’t increase its weight. Ok, so maybe caloric is a fluid that doesn’t have any weight? A rather unusual physical substance…

Here’s another puzzle: you can make two cold things hotter by rubbing them together, i.e. friction generates heat. How? Where does the caloric fluid come from? Well, perhaps caloric can be trapped between atoms and rubbing something releases its caloric, so it is now hot? Ok, but if that’s true, then there will only be a finite amount of caloric fluid trapped between the atoms, so eventually it will run out and the thing you rub will no longer get hot. Scientists tested this. It’s not true – friction never stops generating heat. So is caloric fluid not only weightless but also infinite? A very, very unusual physical substance…

In 1798, Benjamin Thompson, Count Rumford, suggested a different theory: heat is the motion of micro-particles (molecules, atoms). Over time, with other scientific developments on the movement of molecules and atoms, this became accepted. Heat is the kinetic energy of molecules that can be passed from one thing to another.

The theory of caloric fluid turned out be very mistaken. Heat isn’t a kind of fluid at all, but something quite different. So we shouldn’t say that actually, we have reduced caloric fluid to kinetic molecular energy, just changing the meaning of caloric fluid along the way. Instead, we have eliminated caloric fluid – there is no such thing – and explained the phenomena of heat in different terms. Churchland says, ‘the nonexistence of something [e.g. caloric fluid] is established as highly probably… through the acceptance of an explanatorily powerful framework that has no place for it’.

Complexity: Genes and mental states

The reduction of heat to kinetic energy and the elimination of caloric fluid was fairly straightforward. It is unlikely that finding a neuroscientific explanation of thought or consciousness will be anything like as simple. It is worth bearing in mind, then, that scientific reductions can be very messy. A reductive explanation doesn’t have to identify one macrolevel thing with one microlevel thing to succeed.

Genes provide a good example. Genes are the fundamental ‘units of heredity’ that give rise to the observable characteristics of living things. Biologists talked about genes before they knew about DNA. But now we are all told that our genes are ‘in’ our DNA. However, a gene is not necessarily single stretch of DNA (although genes are often misleadingly thought of this way). What we think of as a ‘single’ gene, relating to a characteristic that is inherited from one generation to the next, can involve many distinct segments of DNA (called ‘exons’). It even turns out that the same DNA segment can contribute to different observable characteristics, depending on the stage of development and the environment of the cell in which the DNA is located. Should we say that a gene, or part of a gene, ‘for’ one trait can simultaneously be the gene ‘for’ a different trait or that the same strand of DNA is part of two different genes? Despite all this complexity, biologists accept that DNA, its structure and how it interacts with other things, provides a *reductive* explanation of genes. We can trace a line of cause and effect from DNA sequences to bodily traits, and no one thinks of genes are something *in addition* to DNA.

Understanding reductive explanations in science helps us understand what reductive explanations of mental properties in terms of neurological properties may involve. They may be very messy and complex. There may not be just *one* physical property that we can identify with a particular mental property, but this doesn’t mean we can’t reduce the mental property. The important point is that we can explain mental properties, such as Intentionality and consciousness, in terms of physical properties.

**Elimination and mental properties**

Unlike genes, but like caloric fluid, not all mental properties may survive the process of reductive explanation, claims Patricia Churchland. The way we think about the mind now may be completely changed as neuroscience progresses. But we can’t predict how. Reduction can threaten elimination.

Importantly for eliminative materialism, we aren’t going to get reductive explanations of the mind just working from our everyday psychological concepts of ‘belief’, ‘desire’, ‘emotion’, and so on. These concepts are part of a theory about human behaviour (more on this below). Compared to the workings of the brain, human behaviour occurs over long periods of time (seconds, minutes, days), involves huge complexity and a far wider range of things in space, including other people. A theory of how the mind works can’t reduce a theory of human behaviour to the very fast and tiny processes of neuroscience. An intermediate theory will be needed, e.g. how people process information, what happens when one desire conflicts with another, what processes are involved in a single emotion, how does imagination work, etc.

A good part of this intermediate theory will be developed by cognitive science. Before we can reduce mental properties to neurophysiological properties, we need a much better scientific theory of how the mind works. This will develop *side-by-side* with neuroscience. Only after cognitive psychology and neuroscience have ‘co-evolved’ will reductive explanations be possible. By this point, we can expect that our usual categories for thinking about how the mind works – beliefs, desires, emotions – will have changed and neuroscientific reduction will change them further. We have already rejected many psychological theories from the past, and we can expect this to continue.

For example, it turns out that some people are more easily addicted to substances (food, alcohol, smoking, drugs) than others. We might, common-sensically, say that they have less ‘will-power’. But it turns out that they have different dopamine systems (dopamine is a neurochemical that relates to motivation and a sense of pleasure or ‘reward’ when you get what you want). So now what should we think about ‘will-power’? What is it? Is there really anything such thing?

Paul Churchland on why ‘folk psychology’ might be false

We mentioned above that we have a common-sense theory about why people behave as they do. For example, if someone is thirsty, they will – under normal conditions – look for something to drink. If someone believes it is raining outside, and doesn’t want to get wet, they will – under normal conditions – pick up an umbrella or other covering to keep them dry. And so on. With claims like these, we are able to understand, explain and sometimes predict each other’s behaviour very successfully. We do this by referring to each other’s beliefs, desires, emotions, intentions and so on. Call this body of knowledge ‘folk psychology’.

According to Paul and Patricia Churchland, folk psychology is an empirical theory. As such, it may turn out false, and the central concepts that it uses may, like ‘caloric fluid’, turn out not to refer to anything that exists. So far, however, the claim has only been that this *may* happen. In ‘Eliminative materialism and the propositional attitudes’, Paul Churchland argues that there are three good reasons to think that it *will* happen.

1. There are many aspects of mental life that folk psychology cannot explain, such as mental illness, the nature of intelligence, sleep, perception and learning. Explanations of these phenomena will need concepts that folk psychology lacks.

2. If we look at the history of folk psychology, it reveals no progress since the ancient Greek authors, 2,500 years ago. By contrast, neuroscientific explanations are constantly growing in scope and power.

3. We cannot make folk psychology coherent with other successful scientific theories. In particular, the central idea of Intentionality is highly problematic.

This third objection requires some unpacking. Why think that Intentionality can’t be reduced by other scientific theories, but has to be eliminated? The reason is that it is very puzzling how anything physical could have Intentionality. Intentionality names the property in virtue of which thoughts are directed onto objects or propositions. Thoughts are ‘about’ something, objects or events in the world. For example, I might have a belief *about Paris*, a desire *for chocolate*, be angry *at the government*, or intend *to go to the pub*. In all these cases, my state of mind is ‘directed’ towards an ‘object’, the thing I’m thinking about (Paris, chocolate, the government, going to the pub). However, it seems that physical things are never ‘about’ anything. A particular molecular structure or physical process, described in physical terms, is not ‘about’ anything. For example, digestion is a chemical process, in which acids in your stomach break down food. What is that process ‘about’, what does it represent? Nothing – the question itself is puzzling! But the states and processes of your brain are just chemical states and processes, just like the states and processes of your stomach. So how could brain processes or states ever be about anything? So how could Intentional mental states be states of your brain?

Churchland concludes that folk psychology, with its explanations in terms of Intentional mental states like beliefs and desires, does not fit in with empirically robust theories, such as neuroscience, and so we have reason to abandon it.

**3.7. Objections to Eliminativism**

**Obj: Our certainty about the existence of our mental states takes priority over other considerations**

We can object that eliminativism is simply very counter-intuitive. What could be more certain – indeed, what could be more immediately and directly obvious – than that we have thoughts, desires, emotions, beliefs and so on? Descartes took ‘I think’ to be his first certainty, and for good reason. Nothing, it seems, could be more certain to me than the fact that I have mental states. So no argument could be strong enough to justify giving up such a belief.

But appeals to what is obvious are problematic in the history of ideas. For instance, isn’t it just obvious that the sun moves round the Earth? Just look. And yet it is false. Descartes took it as obvious that there can be no thoughts without a thinker, so he was certain that he was a thinking substance. And yet there are good reasons to believe that there are no substances whose essence it is to think, and many philosophers have argued, along with Buddhists, that there is no ‘self’. Similarly, my ‘mental states’ may not be what they appear to be.

More significantly, the objection misunderstands the Churchlands’ claim. They do not deny the existence of psychological phenomena as such. They accept that the phenomena that we conceptualise as ‘thinking’ occur or again that we experience pain; they deny that folk psychology is the correct theory of their nature. Thinking is not defined by Intentionality as folk psychology understands it, and pain is not a matter of qualia. Instead, they claim that neuroscience will provide the correct account of what these are. As a result, there will be a revolution in our mental concepts. But we won’t cease to feel pain just because we understand what it is in neurophysiological terms. While this revolution is difficult to predict, Paul Churchland argues that explanation will have no place for concepts like ‘Intentionality’, and whether we understand ‘consciousness’ as we do now is also something we may doubt.

All we can be ‘certain’ of is the existence of the phenomena we want to explain. But, the Churchlands argue, appealing to beliefs and desires, Intentionality and consciousness, is not appealing to the phenomena, but to a particular explanation or understanding of them. These concepts are all part of a theory, folk psychology, and we should reject these concepts if the theory that replaces folk psychology has no place for them. We can’t reject unorthodox new ideas just because they are unorthodox.

**Obj: Folk psychology has good predictive and explanatory power (and so is the best hypothesis)**

Paul Churchland criticises folk psychology for its explanatory failures concerning mental illness, sleep, learning, etc. But we can object that this is unfair. Folk psychology is not intended to be a theory of these aspects of mental life, so it is no criticism that it does not explain them. It is only meant to explain human behaviour; or even more specifically, human action. Here, it is incredibly successful. If I know what you want and what you believe, I can predict whether you’ll study hard for your exams. If someone asks me why you went to the cinema last night, I will answer by talking about your love of films and so on. By contrast, neuroscience is almost useless at predicting whether you’ll study hard for your exams or explaining why you went to the cinema last night.

Furthermore, folk psychology is the basis of developments in psychology that have extended its predictive and explanatory power. For instance, ideas about unconscious beliefs and desires have become part of folk psychology. The Greeks used an idea of fixed and unchanging ‘character’, whereas now we tend to appeal more to the situation someone finds themselves in. The importance of situation is a finding in recent empirical psychology, and there are many such findings and theories that use folk psychological concepts and ideas. To eliminate the concepts of beliefs, desires, and other Intentional mental states would do away with much scientific psychology as well as folk psychology.

What this shows is that we don’t have good reasons to think that folk psychological concepts, and especially the concept of Intentionality, will be eliminated as neuroscience develops. We can accept the Churchlands’ insistence that we should only retain concepts that are part of the most powerful explanatory theory, but argue that folk psychology is and will continue to be part of such a theory. The hypothesis that we have Intentional mental states remains the best hypothesis for explaining human behaviour, and won’t be replaced in favour of a neuroscientific theory that eliminates Intentionality.

Eliminativism could reply that these objections are not very strong. First, we need to know how human action or behaviour relates to the rest of mental life. To have very different sorts of theories – folk psychology, neuroscience – explaining different aspects of the mind is unsatisfactory. Second, the developments in folk psychology are relatively superficial. Our folk psychological explanations of behaviour are still far less powerful than the kinds of explanations we find elsewhere in the sciences. The only way to address this problem is to look to neuroscience. Finally, the challenge of explaining how physical states and processes can have Intentionality remains.

**Obj: The articulation of eliminative materialism as a theory is self-refuting**

The Churchlands claim that folk psychology and our commonsense mental concepts comprise an empirical theory. This is why we can think about proving that it is false and eliminating its concepts in light of scientific progress. But there is good reason to suppose that they misunderstand folk psychology. We can argue for this indirectly, focusing again just on the case of beliefs and their Intentional content.

Eliminativism presents arguments, which are expressions of beliefs and rely on beliefs about what words mean and how reasoning works, in order to change our beliefs about folk psychology. Yet, if we turn Paul Churchland’s prediction into a solid claim, eliminativism claims that there are no beliefs. But if that is true, what does eliminativism express and what is it trying to change? If there are no beliefs, including no beliefs about meaning, no beliefs linked by reasoning, then arguments for eliminativism are meaningless. An argument for eliminativism refutes itself – it concludes that there are no beliefs but it must presuppose that there are beliefs.

Eliminativists reply that this objection begs the question. It presupposes that the correct theory of meaning and reasoning is the one that folk psychology gives (in terms of Intentionality). Compare the nineteenth-century argument between people who thought that to be alive required some special energy, a ‘vital force’, and those who said there was no such force. The vitalists could argue that if what their opponents said was true, they would all be dead! Yet now we know there is no special ‘vital force’, that life arises from ordinary chemical reactions. Life just is certain processes, not some special property that living things have in addition to these processes. Eliminativism simply claims that we need a new theory of what it means to assert a claim or argument. What meaning is may turn out to be certain neurological processes.

But we can press the objection. Eliminativism predicts that Intentional content will be eliminated. The very ideas of meaning, or ‘making sense’, of ‘true’ v. ‘false’ belief, or ‘reasoning’ itself, are to be rejected, as they all rest on Intentional content. Claims and arguments are all ‘about’ something. This idea can’t be eliminated in favour of some alternative. The analogy with vitalism fails. Anti-vitalists accepted that they needed to be alive to make their claims, but offered an alternative account of what ‘life’ is. Eliminativists claim that they do not need Intentional content to make their claims. Without having some alternative account of meaning which doesn’t use Intentional content, this is what is inconceivable. We cannot conceive that folk psychology is false, because that very idea, ‘folk psychology is false’, presupposes the folk psychological concept of Intentional content. At least until we have another, better theory of meaning, the assertion that eliminativism is true undermines itself.

On this view, folk psychology – or at least, the central concept of Intentionality – turns out not to be an empirical theory (which might or might not be wrong), but a condition of intelligibility, a condition for thinking, reasoning, and making claims at all. So we can’t eliminate it. That means that Intentional mental states and properties must exist. They are therefore either reducible or irreducible to neuroscience. If Paul Churchland is right that we cannot reduce Intentional content to neuroscience, this isn’t an objection to Intentional content. It is an argument in favour of the irreducibility of mental properties.

**4. Functionalism**

Functionalism claims that mental states are ‘functional’ states. We will need to understand what a ‘function’ is in more detail, but as a first definition, functionalism is the view that each mental state consists of a disposition to behave in particular ways and to have certain other mental states, given certain inputs from the senses and certain other mental states. In other words, we can give an analysis of what mental states are in terms of their ‘inputs’ and ‘outputs’. The inputs are inputs from the senses and other mental states; the outputs are behaviour and other mental states. The complete description of the mental state’s outputs for each possible set of inputs, is the description of its function. It describes what the mental state does.

**4.1. What is a function?**

**Causal role functionalism**

Most functionalists understand the relations between inputs, mental states and outputs causally. Any functional state can be described in terms of what typically causes it, and what it typically causes in turn. We will call this ‘causal role’ functionalism.

There are many kinds of functional state – states that fulfil a functional role. For example, in biology, ‘being an eye’ can be understood in terms of functional role. There are lots of different types of eyes that work in different ways and have different physical properties – human eyes, fish eyes, fly eyes, etc. What makes them all eyes is what they do – convert light waves into neural signals to enable an organism to navigate its environment. In biochemistry, ‘being a poison’ is also a functional property. There are lots of different sorts of poisons that work in different ways and are made of different chemicals. But what makes them poisons is their harmful chemical effect on living creatures. In engineering, ‘being a carburettor’ is a functional property. A carburettor is that part of an internal combustion engine that mixes air and fuel. They can be different sizes and shapes and made out of different materials. And there are lots of other examples. In all these cases, we can define what something is – an eye, a poison, a carburettor – in terms of its causal functional role.

Functionalism argues that the same is true of mental states and properties. What it is to be a mental state is just to be a state with certain typical causal relations to stimuli, behaviour and other mental states. Different mental states differ in their typical inputs and outputs. For example, the typical causes and effects of pain are quite different from the typical causes and effects of a belief that snow is white.

**A computational notion of function**

In fact, functionalist theories began in the 1960s with a different notion of function, related to computers. On this view, the mind essentially works like a computer. Mental states can be compared to software – the instructions for how the machine operates.

In ‘The nature of mental states’, Hilary Putnam first explained what he meant by ‘function’ in terms of how simple computers work. The functioning of a computer can be described by a ‘machine table’. As Ned Block explains in ‘Troubles with functionalism’, this is a long list of conditional statements of the form ‘if the machine is in state S1 and receives input I1, then it produces output O1 and goes into state S2’, ‘if the machine is in state S1 and receive input I2, then it produces O2 and goes into state S3’ and so on. For example, a drinks dispenser that sells drinks at 70p would have a machine table that includes ‘if the machine is in state S1 and receives input of 20p, it should output the message ‘Insert 50p’ and go into state S2’; ‘if the machine is in state S2 and receives input of 50p, it should output a drink and go into state S1’. The machine table lists every possible combination of state and input, and assigns each combination an output.

Machine tables describe the operations of software. And software can be implemented by different systems. For instance, Microsoft Word is a programme that runs on desktop computers, tablets and phones. These machines have different physical constructions, different hardware. But that doesn’t matter, says functionalism. The ‘states’ referred to in machine tables are defined just in terms of inputs, outputs and other states. All that matters, then, is that the hardware – whatever it is – can perform the functions that the machine table describes.

Putnam claims that mental states are simply machine table states. Any mental state, such as being in pain or believing that Paris is the capital of France can be completely described by a set of states and range of inputs within the machine table. Anything that can receive those inputs and have the functional states described by the machine table has the relevant mental state.

**4.2. Functionalism and multiple realisability**

**Inner states**

Whether it understands functions in terms of causal roles or in terms of machine tables, functionalism claims that for something to have functional states, it must have a complex internal organisation. If a functional state is a state with a particular causal role, that causal role will need to be filled by an inner state of whatever possesses the function. For example, to fulfil its function, an eye has to have parts that enable it to convert light waves into nerve firings. Different types of eye have different parts, different structures, but they must all have some structure or they couldn’t enable the creature to see. Or again, a machine that implements a machine table must have a number of distinct physical states that it moves between in response to various inputs and that produce distinct outputs. Again, we need not know what these inner states are, what they are made of, or exactly what mechanisms make them work as they do, but there must be inner states that match each of the functions described by the machine table.

This applies just as much to mental states, since mental states are functional state. For something – whether it is a machine or an animal or a human being – to have mental states, it must have a complex organisation of inner states that work in ways that fulfil the necessary functional roles. These inner states could be states of the brain, but they don’t have to be. Things without brains could have mental states, as long as the relevant functions are performed by some part of them.

Functionalists say that the inner state ‘realises’ the function – it has that functional property. Using our earlier examples, for each eye, some arrangement of light-sensitive and other cells realises the functional property of being an eye; for each poison, some chemical state or other realises ‘being a poison’. In each instance, the causal role that defines what it is to be an eye or a poison is played by some biological or biochemical state or other. What this is can vary from one case to another. The state will be whatever state fulfils the functional role. The nature of the inner state that realises the function isn’t important.

The functionalist argues that each mental property, e.g. ‘being in pain’, is also a functional property. There may be lots of different states, e.g. different brain states, that have this functional property. The states can vary from one species to another. But as long as some state of the creature has the function that defines pain – given certain inputs, it causes certain outputs – then the creature is in pain.

**Multiple realisability**

Putnam argues that mental properties are not identical to physical properties because the same mental property can be related to or supervene on different physical properties. For example, the brain states that relate to pain may well be different in different species, in humans and birds, say, but pain is the same mental state. If this is true, there are creatures who, when they are in pain, have different physical properties from us when we are in pain. Therefore, ‘being in pain’ cannot be exactly the same thing as having a particular physical property. This is the argument from ‘multiple realisability’.

The term ‘realise’ here means ‘to give actual form to’ or ‘to bring into reality’. For instance, we can talk of a design for a dress being beautifully realised in the final product. Similarly, in metaphysics, philosophers talk of one property ‘realising’ another. To say that a particular neurophysiological property ‘realises’ pain in human beings is to say what pain is, the form of existence it has, is given by that neurophysiological property.

According to functionalism, mental properties are multiply realisable because functional properties in general are multiply realisable. As we’ve seen, ‘being an eye’ is multiply realisable. What identifies the property ‘being an eye’ is a particular causal role. In humans, in fish, in flies, the occurrence of a particular arrangement of cells fulfils this causal role, and so has the functional property of being an eye.

Functionalism identifies mental properties not with the physical properties of brain states, but with what brain states can do. And what one brain state can do may be something that a different brain state, or even a state of something that isn’t a brain, e.g. a computer, can also do. Things with very different states – different constitutions or internal organisation – can realise the same mental states as long as they are states with the same causal roles (or realise the same machine table). The nature of the state – biological, electronic, etc. – doesn’t tell us anything essential about the mental state, which is purely a matter of functional role.

This also explains why functionalism is compatible with both dualism and physicalism. The metaphysical nature of the state that plays the functional role could be anything. Mental states are mental states in virtue of what they do, not in virtue of the nature of the substances or properties that realise those mental states. Mental states could be realised by physical states, e.g. of the brain, or they could be realised by states in a distinct mental substance, or they could be realised by a creature composed of both mental and physical substances. However, while functionalism is compatible with dualism, most functionalists are physicalists.

Functionalism is compatible with physicalism because functional properties in general supervene on physical properties. As we’ve seen, functional properties occur throughout science, e.g. being an eye. Once something has a certain internal complexity and organisation, and it can receive certain inputs and produce certain outputs, then it can have the functional property of being an eye. All this – its internal structure, inputs and outputs – can be described and explained in terms of its physical properties. Nothing more is needed. Its physical properties fix its functional properties. These functional properties are not themselves physical properties, according to functionalism, because there are lots of different ways in which eyes can be constituted physically. However, functional properties are properties which are realised by physical properties operating in causal relationships. And what is true of ‘being an eye’, functionalism claims, is true of ‘being a pain’ or ‘being a belief’. If physicalism is true, then it is a physical substance and its physical states, e.g. the physical states of a brain, that realise mental states.

**4.3. The challenge to functionalism**

Phenomenal properties are properties which give an experience its distinctive experiential quality, ‘what it is like’ to undergo that experience. According to some philosophers, phenomenal properties are intrinsic and non-Intentional. If they are, then they are ‘qualia’. (For more on qualia, see the handout ‘What do we mean by mind?’.)

The objection to functionalism is this: if phenomenal properties are qualia, then they cannot be completely understood in terms of their causal roles (or inputs and outputs on a machine table), because these are relational properties, not intrinsic properties. It is not what causes them and what they cause in turn that makes pain or the smell of coffee or the visual sensation of red what it is. What it is like to experience these mental states – how pain feels, how red looks, how coffee smells – can’t be analysed in terms of functions. Yes, of course, how pain feels is important to what it causes, e.g. it causes you to cry out or withdraw your hand from the fire. But the feeling of the pain isn’t just these causal relations. So functionalism can’t explain phenomenal properties.

P1. Qualia, by definition, are intrinsic, non-Intentional properties of conscious mental states.

P2. Intrinsic, non-Intentional properties cannot, by definition, be completely analysed in terms of their causal roles (or as machine table states).

C1. Therefore, if qualia exist, some mental properties cannot be analysed in terms of their causal roles (or as machine table states).

P3. Functionalism claims that all mental properties are functional properties which can be completely analysed in terms of their causal roles (or as machine table states).

C2. Therefore, if qualia exist, functionalism is false.

P4. Qualia exist.

C3. Therefore, functionalism is false.

The controversial premise is (P4). In what follows, we will look at a thought experiment that tries to establish this premise.

Thought experiments are designed to test a hypothesis or philosophical claim through imagining a hypothetical situation, and coming to a judgment. In philosophy of mind, the most common kind of thought experiment is one in which we are asked to judge whether the hypothetical situation is possible or not. Thought experiments are used when actual experiments are either practically or physically impossible, or when the judgments concern matters that cannot be investigated by empirical means, e.g. the correct application of concepts or metaphysical questions of identity.

The possibility of a functional duplicate with different qualia (inverted qualia)

We can show that phenomenal properties cannot be understood just in terms of their functions if we can show that it is possible for two people to have states with identical functions but different phenomenal properties. The most popular version of this objection is known as the case of ‘inverted qualia’.

Suppose that you and I are looking at ripe tomatoes and fresh grass. Because we have grown up in the same linguistic community, we have learned to use the word ‘red’ to describe the tomatoes and ‘green’ to describe the grass. So we both say that the tomatoes are red, the grass is green. But the particular way that tomatoes seem to me is the way that grass looks to you, and vice versa. Functionally, we are identical, and yet we have different colour experiences. ‘The way grass looks to you’ and ‘the way grass looks to me’ are functionally identical; both are caused by the same inputs (grass) and cause the same outputs (e.g. saying ‘grass is green’). But they are not identical in terms of their intrinsic properties. They refer to different qualia.

Of course, we might not know whether this is true or not. But that is irrelevant. The objection is that inverted qualia are possible. If functionalism were true, inverted qualia would be impossible. So functionalism is false.

**Patricia Churchland on inverted qualia**

The functionalist can reply that in the case described, you and I are not, in fact, functionally identical. There are going to be small, but very important, differences, because the causal relations of phenomenal properties are very complex. In *Brainwise*, Patricia Churchland argues that we have no good reason to think that qualia can be inverted in the way the thought experiment describes.

In presenting her argument, Churchland is defending eliminative materialism rather than functionalism. But because she argues that phenomenal properties are not intrinsic properties, her defence works for both theories.

She starts by making the objection from inverted qualia clearer. The main claim in the objection is that you and I – or our brains – could function in exactly the same way, but we would have different qualia. This is not being proposed as an empirical hypothesis, e.g. that you and I really do see red and green differently. Why not? Because empirical hypotheses are tested against the evidence. First, we have no evidence from neuroscience that identical brain functioning gives rise to different conscious experiences in different people. Second, as an empirical hypothesis, it is poor, since it proposes that there could be empirical differences (in our conscious experience) that are undetectable, since they make no functional difference. But science does not proceed by supposing undetectable facts! So if the inverted qualia objection were empirical, then it is either false or bad science.

So in saying that you and I could function in exactly the same way, but have different qualia, ‘could’ must mean ‘it is conceivable’. So the argument is like other arguments from conceivability, and faces the same objections as other arguments from conceivability. We can question whether it is conceivable; whether, even if it is conceivable, it is possible; and whether its possibility tells us anything about whether there actually are qualia (i.e. whether phenomenal properties are intrinsic and non-Intentional).

Churchland then points out that the thought experiment is much too simple. First, every colour that we can discriminate has unique similarity and dissimilarity relations to all surrounding colours. For instance, red is more similar to orange than green is, while green is more similar to blue than red is. So we can’t simply switch red and green without messing this up. If you and I saw red and green ‘switched’, then we wouldn’t agree on whether red was more similar to orange or blue. And this is a functional difference.

One response would be to change the thought experiment – it is not just red and green that are inverted, but the whole spectrum. This could keep all the similarity relations as well. Someone who sees red as green also sees orange as blue, and so they say that red is similar to orange, but what they see is what I see when I look at green and blue.

But this meets another problem. Human beings can make much finer discriminations in green, yellow and orange than we can in blue. If we inverted everything, this would be apparent from behaviour, as whoever sees the inverted colours would be able to make finer discriminations among blue than the rest of us can, and fewer discriminations in green, yellow and orange. And so it is empirically impossible for someone to have inverted qualia without functional differences.

But isn’t it still conceivable that we could correct for this as well? The person with inverted qualia has the same objective discrimination abilities as the rest of us, even if subjectively, they are discriminating between blue qualia in ways that we can’t. This is conceivable, but, asks Churchland, should we say that this person really sees the same colours as us? The colours that they see bear new similarity relations to every other colour, and they have different powers of discrimination. What is essential to a colour being the colour it is?

The qualia theorist argues that phenomenal properties are intrinsic – they are essentially what it is like to experience them. In other words, it is conscious introspection that identifies whether two colours are the same or not. But is this right?

We can offer an explanation of our experience of colour in terms of our physical constitution and how it functions. For example, why does colour have the three dimensions of red, green and blue? The answer has to do with types of colour-sensitive cells, ‘cones’, in our retinas and the way they are wired up to the brain. If we say that our experience of qualia is the way the brain processes this information, we can explain our colour experience – similarities and dissimilarities, what colours we can see and what we can’t etc. But if qualia are something distinct, we have no explanation for why we see colour as we do. What colour experience is should be decided by who has the best explanatory theory of colour experience, and not by thought experiments.

**Block on the possibility of a functional duplicate with no qualia**

The second objection to a functional analysis of consciousness tries to show that phenomenal properties cannot be understood just in terms of their functions because it is possible for two systems to have states with identical functions but one system (you, say) has phenomenal properties, but the other does not. The most popular version of this objection is Ned Block’s China thought experiment.

In ‘Troubles with functionalism’, Block frames his argument in terms of machine-table functionalism. Suppose we have a complete functional description of your mental states. For each and every one of your mental states, we have an input–output analysis, giving us a machine table for your mind. Now imagine that a human body, like yours, is connected up via its sensory and motor nerves not to a brain but, through electronic transmitters, to the whole population of China. The Chinese are linked up to each other by two-way radios, and some of these are linked up to the input and output nerves of the body. (Block picks China because the population of China is 1 billion, which may be enough to fulfil the functions that comprise your mental states). Then, for a short time, the Chinese population realizes the same machine table that describes the functions of your mental states.

According to functionalism, this should create a mind; but even if we could accept that this set-up could have Intentional mental states such as beliefs or thoughts, it is especially difficult to believe that there would be a ‘Chinese consciousness’. If the Chinese system replicated the functioning of my brain when I feel pain, would something be in pain? What? Is there something it is like to be this system? The objection is that the Chinese system, although it duplicates your functioning, can’t duplicate your mind, because some mental states are qualia, and the system can’t have qualia because they are not functional states.

Functionalists can reply that the Chinese system won’t be functionally identical to you. For instance, it could be disrupted by things that your mind isn’t disrupted by, e.g. the radios running out of batteries or the system being disrupted by bad weather.

True, but irrelevant, says Block. First, although this *could* happen, if it doesn’t, then we have functional duplication, and the functionalist must say that the Chinese system is conscious. Second, these disruptions don’t count as inputs or outputs, any more than having a brain tumour counts as an ‘input’ to our mental states. It is not part of their functioning – that’s why they are disruptions.

Functionalists can object that the Chinese system is much slower than our brains. But, replies Block, why should this matter for whether it has mental states? Couldn’t there be much slower minds than ours? In any case, this is just an objection about what is physically possible. A Chinese system that operated as fast as our brains is still metaphysically possible.

**A Physicalist Response**

If Block’s objection works, then not *everything* about the mind can be explained in terms of functions. But perhaps we can combine functionalism and the type identity theory to argue as follows.

If the Chinese system can have Intentional mental states, then functionalism provides an accurate account of all mental states except for consciousness of phenomenal properties, which involves qualia. Why should this be? We could argue that the intrinsic properties of qualia depend on the specific *physical* properties of the system that realises the functional states. We saw this type of argument in Patricia Churchland’s response to the objection from inverted qualia above. How colours look to us isn’t (just) a matter of what causes the colour experience and what effects it has, it also depends on our physiology – the types of cones we have and the way our brains are wired. So what mental states something has depend on its functional properties *and* its intrinsic physical properties. Mental states are still nothing more than physical states playing a functional role. A physical, functional duplicate of a person with consciousness will have the same conscious stat