# PHILOSOPHY OF MIND, BRAIN AND BEHAVIOUR

Marc Slors, Leon de Bruin & Derek Strijbos

Boom | Amsterdam

# 1

# The mind-body problem

The mind-body problem is the most central puzzle in the philosophy of mind, brain and behaviour. All problems and debates presented in the rest of this book are connected with this problem and with the various attempts to solve it that are discussed in this first chapter. The term 'mind' will be used here as an umbrella-term. It does not necessarily refer to one thing. Rather, it refers to a range of processes and states that we commonly refer to as 'mental', such as thinking, having beliefs and desires, perceiving, being conscious, experiencing emotions, and forming and acting on intentions. How do such processes and states relate to the physical states of our bodies and brains?

Do mental states and processes belong to a different realm of reality than the physical realm studied by the natural sciences? Many people think so, even today. Most scientists and philosophers, however, think that attempting to explain the mind as a set of natural, physical phenomena is the only way forward to understanding what these phenomena are, how they can exist and how they can shape our behaviour and our lives. In the next section we will discuss some of the reasons for this. Can we indeed explain the mind as a natural phenomenon without reference to immaterial souls or other non-natural phenomena? If so, how? This is the mind-body problem.

Mental states and processes have a number of properties that make it very difficult to understand them as natural phenomena that are part of the physical world. For example, thoughts, beliefs, hopes, fears and dreams all have the strange property of be-

ing *about* something. You can think about your next meal, about Caesar crossing the Rubicon or about black holes, and an infinite number of other things. None of these topics – your actual next meal, the historical event of Caesar crossing the Rubicon or real black holes – are themselves part of your thoughts. Your thoughts only *refer* to them. How can such reference – such 'aboutness' – be physical? (See Chapter 3 for more on this problem).

Another example of properties of mental states that are hard to fit into the physical world of the natural sciences can be found in experiences such as tasting the sweetness of sugar, seeing the colour red, or feeling an itch. Such experiences have a very specific experiential quality. There is 'something it is like' to have them. This 'something' may well be impossible to describe. You need to experience it to know it (try to explain what sugar tastes like to someone without taste buds). Such experiences are deeply and fundamentally subjective. Just as in the case of the 'aboutness' of thoughts and beliefs, the point here too is that we find no easy parallel in other physical phenomena. No other physical states have these ineffable subjective qualities. Indeed: how can subjective qualities even be part of an objective physical worldview? (See Chapter 2 for more on this problem).

Properties of mental states and processes such as these make it tempting to think that the mind must be a non-physical entity. But, as will become clear in the next sections, this idea is very problematic, both from a scientific perspective and from a logical, philosophical perspective. Thus we are left with the task of explaining (or explaining away – see section 7) seemingly non-physical properties of mental states and processes in terms that are compatible with the physical world of the natural sciences. In the course of the twentieth century a number of theories have been proposed with exactly this aim. In this chapter we will discuss the seven most influential ones.

Before we can discuss these theories however, it is necessary to go back in time a little bit further. All contemporary theories of the mind can be understood as responses to – or better, motivated rejections of – the idea that the mind resides in an immaterial

soul that is distinguishable from the body. The most consistent and influential version of this idea is René Descartes' (1596-1650) so-called substance dualism. All contemporary solutions to the mind-body problem are attempts at abandoning the Cartesian way of thinking about the mind. Yet, as we will see, many contemporary theories of the mind still have some distinctly Cartesian features.

## 1.1 Substance dualism

Substance dualism is the theory according to which people consist of two parts, a material body and an immaterial soul. The soul is held to be the seat of our mental states and processes. Body and soul are thought to be able to exist separately. In philosophical jargon: they are separate substances. The term 'substance' means 'something that can exist independently' – thus, a chair is a substance, but the leg of a chair is not, because it needs the rest of the chair to exist as a chair's leg. Claiming that the soul and the body are distinct substances need not imply that the soul and the body are *actually* separated. On the contrary, substance dualists emphasize that there is a continuous interaction between the soul and the body. Perception, for instance, starts with bodily processes – sensory impingements – but may result in conscious experience. Conversely, the will is thought to reside in the soul but its effects usually consist of bodily actions. In daily life, body and soul form one unity – man, but in principle, body and soul can exist as separate entities.

Substance dualism is often associated with religion, life after death, spiritism and paranormal phenomena. René Descartes' version of this theory, however, has little to do with all that. Descartes did not think, for instance, that the soul has a specific ghostly shape or that it is made of the kind of immaterial stuff that nineteenth-century spiritists dubbed 'ectoplasm'. According to him the soul does not even occupy space – it has, in philosophical jargon, no 'extension'. Only material things have extension, according to Descartes. In fact, he thought that this is exactly what defines matter. The essence of the immaterial soul, on the other hand, is what

he called 'thinking'. This term should not be taken too narrowly. In Descartes' usage it encompasses not only intellectual mental processes but all conscious processes. Thus, imagining, perceiving, desiring, believing, doubting, hoping, and dreaming are all characteristics of the soul.

### *1.1.1 Doubt as argument*

Descartes' substance dualism is the product of his quest for the foundations of scientific knowledge. Such foundations, he assumed, should consist of knowledge that we cannot doubt. How can we find such knowledge? Well, simply by doubting everything we presume we know, in the hope that we come across a piece of knowledge we cannot possibly doubt. It is crucial here that Descartes' doubt was radical. It does not matter if it is reasonable to doubt some presumed knowledge, all that matters is if we can conceivably doubt it. Very little of what we think we know is resistant to such radical doubt. We can conceivably doubt the existence of the outside world, for instance. We can even doubt the existence of our own bodies. For all we know we may be in a scenario much like that in the movie *The Matrix*, in which our minds are in fact computer programmes that provide us with the illusion of having a body. That is, according to Descartes it is logically possible (though obviously very improbable) that all that exists is our *experience* of the world and our bodies, but not these things themselves. In fact, he argued, there is only one thing that we cannot doubt, when we doubt so radically, namely that there is something or someone – me – who is currently doubting. Doubt is a form of thought. Hence, the foundation upon which Descartes builds his worldview is this: 'I think (I doubt) therefore I exist', or, in Latin *cogito ergo sum*. The essence of this 'me' is that it is a 'thinking thing'.

In his *cogito* argument, Descartes was not looking primarily for a proof of substance dualism. His reasons for accepting dualism are based on the assumption that material objects cannot think. Material objects cannot reason and cannot be conscious according to him – we will come back to this below. Others, however, do see a direct proof of dualism in the *cogito* argument. In order to see why,

we need to invoke a principle that was first proposed by Gottfried Leibniz (1646-1716). Leibniz held that we can only say that *x* is really the same thing as *y*, if *x* and *y* share all their properties. In order to illustrate this principle of the 'identity of indiscernibles', take the example of two identical billiard balls: both balls are equally big, made of the same material, equally smooth, and have the same colour, etc. Even though they are identical in all their properties, they are still two billiard balls (or in philosophers' jargon: they are qualitatively identical but not numerically identical). This is because there is one property that they do not share: their place in space and time. If they would share that property too they would not be two balls but only one.

Leibniz' principle can be used to argue that Descartes' thinking 'me' cannot be the same entity as a material body or brain (or so it seems – see 1.1.3). For Descartes argued that we can conceivably doubt the existence of our own bodies and brains, but we cannot doubt our own existence as thinking (doubting) things. Hence, these thinking things – us – appear to have a property that our bodies and brains lack. And by Leibniz' principle, this would mean that these thinking things must be entities that are distinct from our brains and bodies.

### *1.1.2 Two features of Cartesian dualism*

Descartes' thinking 'me' is his conception of the soul. It is an immaterial entity. As said, this entity interacts with the material body. The immateriality of the soul, and hence its existence as distinct from the material body, is a prominent feature of the Cartesian view of the mind. When philosophers and scientists present themselves as being anti-Cartesian, they usually mean that they are opposing the idea that people consist of two separable parts and that they do not believe in an immaterial soul. For some philosophers and for most scientists, the only meaning of being 'Cartesian' is to believe in an immaterial soul.

There is, however, a second feature of Cartesian dualism that remains intact in many contemporary philosophical and scientific theories of the mind. Descartes conceived of the mind as a domain

of the thinking 'me' that is separated from the outside world. It is connected to the outside world only *indirectly*, through the senses (which provide input for the mind) and through behaviour (which constitute the mind's output). Knowledge of the world resides in the mind, obviously. But it consists of what he calls 'ideas'. Ideas exist in the mind but they are often *about* the world. Thought is the manipulation of ideas. True knowledge consists of ideas that correspond to how the world really is. If we let our behaviour be guided by true knowledge, the mind is able to cause the body to act successfully in the world.

It is easy to combine these two features of Cartesianism – (1) the immateriality of the soul and (2) the separation of the outside world and the inside mind. But the second feature does not necessarily presuppose the first. Many philosophers and scientists who reject the idea of an immaterial soul – and hence consider themselves as being anti-Cartesian – believe that the mind consists of brain processes that are only connected indirectly, through the senses and behaviour, to the outside world. In that sense they are still Cartesian, despite their rejection of immaterial souls. The soul-body dualism is traded for a brain-body dualism.

### *1.1.3 Problems with arguments for dualism*

Descartes' own arguments for dualism were based on the idea that material objects or entities cannot think. In particular, material entities were thought to be incapable of using language, of reasoning and of having conscious experiences. We shall leave conscious experience aside for now; this is a very complex issue to which we will devote the next whole chapter. It suffices to note here that, as we shall explain in Chapter 2, there are many philosophers nowadays who think Descartes was wrong in thinking that material entities cannot be conscious in principle.

Where language and reasoning are concerned, it is easier for us, now, to recognize that Descartes' reasons for accepting dualism were insufficient. As twenty-first-century people we are familiar with computers that can reason and that produce and understand language. This does not mean that these computers use language

and reason in exactly the way we as human persons do. But there are no reasons in principle to deny that one day they might be able to do so, and without such principled reasons, Descartes' reasons to accept dualism fail.

What remains, however, is the argument from doubt. We can doubt the existence of our brains and bodies, but not the fact that there is someone who doubts. Hence, or so the argument goes, this someone cannot be identical with her brain and body. For in order for there to be identity, the thinking entity must share *all* her properties with her brain and body. Does this argument really show that accepting substance dualism is inescapable? In fact it does not. Abstractly put, the counter-argument here is that what someone is entitled to think about *x* (e.g. whether or not they can doubt *x*'s existence) should not be considered a property of *x* in the sense of Leibniz' principle of the identity of indiscernibles.

Consider the following example. At some point in time people thought that the morning star and the evening star were different planets, simply based on the fact that these planets were perceived at different times. But when and how something is perceived is not a property of that object. Hence the fact that people thought of the evening star and the morning start as different planets does not stand in the way of the possibility that, even by Leibniz' principle, they are the same star – which in fact they are: Venus. Just as being perceived at a certain time is not an intrinsic property of a star, so whether or not we can doubt an object's existence is not a property of that object either.

The relevant distinction here is between what philosophers call ontology and epistemology. Ontology is the study of what really exists; epistemology is the study of what we know and how we can know it. The distinction between the mind and the body in terms of what we can doubt is an epistemological distinction. It is not about the mind and the body as they exist, but between the mind and the body as we know them. The idea that the mind and the body are distinct entities, however, is an ontological conclusion. It is about a distinction that exists independently of what we think or know. The argument for substance dualism based on doubt fails

because it draws an ontological conclusion from an epistemological difference.

### *1.1.4 The interaction problem*

The arguments discussed in the previous sub-section show that there are no logically inescapable arguments for dualism, but a lack of arguments for dualism is in itself not an argument against it. However, such arguments do exist. One of the most important arguments stems from what is known as the interaction problem.

The interaction problem was put on the philosophical agenda by a contemporary of Descartes, princess Elisabeth of Bohemia (1618-1680) who corresponded extensively with him. How can an immaterial soul cause a body to move as in when someone moves her arm voluntarily? Conversely, how can the material body cause changes in the immaterial soul, for instance in conscious sensory perception? According to Descartes material substances belong to an entirely different realm of reality – the physical realm of extended objects – than immaterial souls – which belong to the non-extended realm of thinking things. However, if they belong to different realms, how can there be interaction between the two?

Descartes' initial answer to this question involved locating *where* the soul influences the body (he thought this occurs in the epiphysis or the pineal gland, a tiny structure in between both cerebral hemispheres). It also involved contentions about forces streaming from the soul into the body. But to princess Elisabeth's complaint that that is all idle speculation he simply admitted that she was right. However, he argued, the fact that we do not know how the soul and the body interact does not mean that there is no interaction; we know there is. He compared the situation with magnetism. In his time this was an utterly strange and unexplained phenomenon. However, the unavailability of an explanation does not imply that the phenomenon does not exist. And so it is, Descartes held, with mind-body interaction.

Nowadays this defence doesn't hold up. For one thing, the comparison with magnetism fails, for we do currently have extensive theories about its nature and workings. More importantly, we now

have reasons to believe that an explanation of the influence of an immaterial entity on physical processes will never be possible. This is connected with what is known as the 'causal closure of the physical realm.' This term refers to the idea that the occurrence of every physical event has a complete physical explanation. This idea is a point of departure for contemporary physics – it cannot be proven empirically but its likelihood is proportional to the (immense) predictive power of the physical sciences. If the physical world is, indeed, causally closed, as contemporary science has it, all physical events, including events in the pineal gland, have a complete physical explanation. The existence or inexistence of an immaterial soul would make no difference. Otherwise put: the soul would be out of a job when it comes to influencing our bodies and determining our behaviour.

## 1.2 Logical behaviourism

At the end of the nineteenth century, the philosophical probl[em]s pertaining to substance dualism were complemented by m[e]thodological ones. By that time psychology was developing i[nto] an autonomous science. But how can one investigate an i[m]material soul scientifically? Starting from a dualist position, th[e] only possibility for science is to rely on introspection. This [m]eant that subjects were asked to 'look inward' and report o[n] their subjective experiences as accurately as possible. Howe[ver] despite the fact that this method was refined with much s[ci]entific ingenuity by psychologists such as Wilhelm Wundt (1[83]2-1920) and Edward Titchener (1867-1927), serious metho[d]ological problems remained. The most serious problems stem f[rom] the fact that introspective reports can never be verified or f[al]sified by others. Only the subjects themselves have access to th[eir] own experiences, and this makes it hard, or even impossi[ble] to acquire objectively valid scientific knowledge.

In res[pon]se to these methodological problems, John Watson (1878-19[58]) recalibrated psychology as a science of behaviour. The so[-c]alled mind, consciousness and all related notions were banned [by] him as being unscientific. The initial success of this psycholog-

wish to be impolite. In such a case, too, you will pick up the gla and drink its contents. Or you might be suicidal and believ glass contains poison. Again, you will pick up the glass an k from it. Thus, mental holism undermines the one-to-on nection between mental states and behavioural disposition logical behaviourism presupposes.

A third problem with logical behaviourism pe s to the idea that the connection between mind and behavio not causal but conceptual. This is concluded from the fac we can make a distinction between intelligent and non-i gent behaviour before we attempt to back this explanati by a causal explanation according to which intelligent be our is caused by a mind. Ryle writes as if a conceptual con on between mind and behaviour excludes a causal conne . But is that correct? Take as an example the American com ial slogan for a breakfast cereal 'Wheaties is the breakfast hampions!' This slogan certainly makes a conceptual clai is claimed that it is part of the very concept of Wheaties t t is the breakfast that champions eat. However (says Jerry or – see sections 1.5 and 1.7.2) there is also a causal claim hid n the slogan: Wheaties *makes* you a champion; champions hampions *because* they eat Wheaties.

A concep connection, then, need not rule out a causal connection. T ct that there is a conceptual connection between mind a haviour – the fact that we are already able to tell intelli ehaviour from non-intelligent behaviour – does not precl e fact that intelligent behaviour may have a different cause non-intelligent behaviour, and the causes of intelligent behav- r may well be referred to as 'the mind'.

## 1.3 The identity theory

While Ryle in Britain attacked the conceptual dualism between the mind as an inner realm and behaviour as a mere 'outside' affair, a different rejection of substance dualism emerged in the 1950's in Australia (referred to by British logical behaviourists as 'the Australian heresy'). This rejection is not aimed at the inner-outer dis-

tinction but purely at the idea of the mind as being located in an immaterial soul. This so-called 'identity theory' did not find its origins in philosophy, but in experimental psychology. Already in 1933, the psychologist Edwin Boring (1886-1968) suggested that consciousness can best be understood as a brain process. Differently put, consciousness may be *identical* to a brain process (hence the name identity theory).

This suggestion is surprisingly modern, but Boring's contemporaries were all psychological behaviourists. Consciousness was taboo, and Boring's ideas were virtually ignored until the Australian philosopher Ullin Place (1924-2000) rediscovered Boring in the early 1950's. Like many philosophers in that time, Place was inclined to think there is a conceptual connection between mental states and behaviour, more or less as suggested by Ryle. But he was also convinced that logical behaviourism fails as an explanation of consciousness – the pain examples from the previous sub-section illustrate the problem here. Boring convinced Place that as far as consciousness is concerned, the identity theory is the best alternative for the scientifically indefensible position of substance dualism.

### *1.3.1 Consciousness as a brain process*

Place did not merely copy Boring. As a philosopher he was primarily concerned with explaining what we *mean* when we say that consciousness is a brain process. What does the word 'is' – expressing identity – mean here? Take three sentences: (1) 'George Orwell is Eric Blair', (2) 'George Orwell is famous', and (3) 'George Orwell is the writer of *1984*.' All sentences contain the word 'is', but in all sentences the word is used differently. In the last sentence, the word 'is' is used to characterise someone, i.e. to refer to one of his features. A similar usage of 'is' is in definitions such as 'a bachelor is an unmarried man.' In the second sentence, the word 'is' is used to ascribe a property. Both, related, usages of the word 'is' are *not* applicable to the claim that consciousness is a brain process, Place explains. The identity theory is not meant to characterise consciousness or ascribe a property to it. (For that would make the fact that consciousness is a brain process a conceptual fact – which

would make it impossible that some philosophers ever considered consciousness to be immaterial).

The sense of 'is' that is at play in the identity theory is the sense in which the word is used in the first sentence. It is used to indicate identity: George Orwell is the very same man as Eric Blair. Likewise, consciousness is the very same process as some specific brain process. To say that consciousness is a brain process is to say that consciousness is constituted by a brain process. We can say that the Eiffel tower is constituted by a set of iron beams, but just as we cannot define the Eiffel tower as a set of iron beams, we cannot define consciousness as a brain process. The 'is' of the identity theory stands for constitution, not for definition. Thus, though it is logically possible that the mind is immaterial, according to Place, we have discovered that this is actually not the case.

### *1.3.2 Mentalistic language is 'topic neutral'*

Place's position is adopted and further developed by J.J.C. Smart (1920-2012), a philosophical colleague who was brought to Australia by Place. Smart applies the identity theory not only to consciousness, but to the entire mind. He emphasizes that the identity theory is a theory that is made plausible by scientific research. Brain research shows more and more that the mind is manipulated only if and only insofar as the brain is manipulated. This suggests very strongly that the mind is the brain. But as with Place this is not a conceptual claim for Smart. The discovery that the mind is the brain is comparable with the discovery that water turns out to be $H_2O$. It is conceivable that we would have discovered it to have some other chemical structure – but we didn't. Likewise, it is conceivable, according to Smart, that we would have discovered that Descartes was right and that minds are immaterial souls. But that is not what we found.

Smart's main contribution to the identity theory is his contention that mentalistic language – i.e. the way we speak about ourselves and others in terms of what we think, feel, want, etc. – is 'topic neutral'. What he means by this is that such language does not make any commitment as to the nature of our minds. Speaking

of thoughts or feelings is not to make implicit theoretical claims about the existence of immaterial entities. Mentalistic language is neutral with regard to the material or immaterial nature of our minds. This is important, for it implies that we are not obliged by the fact that we speak of minds to complicate our worldview unnecessarily by introducing unexplainable immaterial entities.

### *1.3.3 Type-identity theory*

The identity theory comes in two flavours: type-identity theory and token-identity theory. The word 'type' stands for 'class' or 'category'. The word 'token' stands for an individual object or entity that belongs in a class or category. My neighbour's dog is a token of the type 'Doberman.' The car opposite my house is a token of the type 'Vauxhall Astra'. And the tree next to my house is a token of the type 'Oak.' We will concentrate on type-identity first.

The type-identity theory is the most rigid, the most severe form of identity theory. It is the theory as defended by J.J.C. Smart. According to this theory every mental state of a specific type is identical to a brain state of a specific type. Thus, believing that Amsterdam is the capital of The Netherlands (which is a mental state *type*, a class or category of beliefs, because there can be very many tokens of it: we believe it, and so do you and very many other people) is identical to a very specific brain-state type. This means that everyone who believes that Amsterdam is the capital of The Netherlands shares a specific neural structure.

If the type-identity theory is true, it would in principle be possible to tell what people are feeling or thinking about by inspecting their brains. If every thought of a specific type is identical to a brain state of a specific type, then knowledge of these type-type identity relations will enable us to read people's minds by determining their brain processes.

### *1.3.4 Multiple realization*

According to the type-identity theory a specific mental state type – say pain – is identical with a very specific brain state type. An important problem for this position is posed by the fact that it

seems that animals with different brains appear to be able to be in the same kind of mental state. In other words, different kinds of brains appear able to realize the same kinds of mental states. This is called multiple realisation. Pain is a good example. It seems that many animals are able to feel pain. Yet a cow's brain, say, is rather different from a human brain. If both humans and cows can feel pain, then type-identity theorists would predict that they must share the brain-state types with which pain is identical. But since human brains and cows' brains are really different, there can be no brain-state type that humans share with cows. If type-identity is correct, then, we must conclude that, on reflection, cows and other animals cannot feel pain. This is counterintuitive at best.

One solution for this problem is to accept that there are many different kinds of pain. We can hypothesise that a cow's pain feels somewhat different from human pain. This may be a somewhat plausible assumption. As far as subjective experiences such as pain go, we may well assume that different animal types have different types of experiences. But what if we start looking at, say, visual perception? Take the perception of a dark square surface. Must we assume that, because dogs, cows, humans, horses and monkeys all have different types of brain, they all perceive a different shape? That seems highly unlikely. The much more reasonable assumption is that we all perceive the same square and that the mental state-type 'perceiving a square' can be realized by different brains in somewhat different ways.

Another example of multiple realisation is neural plasticity. After some cases of brain damage, for example, certain brain functions are performed by different parts of the brain. This type of multiple realisation, too, is at odds with the type-identity theory.

### *1.3.5 Token-identity theory*

Multiple realisation is a serious problem for type-identity theory. Contemporary defenders of this theory cannot but deny that it is a real phenomenon (there are independent reasons to be found for such a denial; see section 3.3.3). For many philosophers and scientists, however, multiple realisation is a given that all theories

should be able to take into account. This is why, from the 1970's onwards, many defenders of the identity theory opted for another version of that theory: token-identity theory.

According to this theory, every token of a mental state-type is identical with a token of a brain state-type. But there need not be type-identity. Thus not all tokens of a mental state-type need be identical with tokens of the same type of brain state. In other words: multiple realisation is deemed to be possible. In fact all that the token identity theory says is that mental states are brain states. It makes no claims about which mental states are which brain states. Thus, when the token-identity theory is true and not type-identity, it is not the case that some future surgeon may read your mind simply by determining the brain state-types you are in.

Of course the token identity theory invites the question of how tokens of different brain state-types can be identical with tokens of the very same mental state-type. We will discuss the best-known answer to this question below while discussing a variant of the token-identity theory, functionalism.

### *1.3.6 Identification without explanation*

One of the big problems of the identity theory, in both forms, is its relative lack of explanatory potential. It is one thing to argue that science is showing us that massively complex electrical activity in the soggy grey matter of our brains is identical to thoughts, sensations, intentions, etc. But that is not at all to explain *why* this is so. On the face of it, thoughts, feelings, sensations, and all other mental activity need not necessarily have anything to do with electricity or brains. That is: while having a sensation there is nothing that suggests, to us subjects, that really all the sensation consists of is brain activity. And identity theorists recognize this. Place and Smart explicitly allowed for the logical possibility that we could have found out that minds are souls (it's just that we found out something else).

Identity theorists tend to respond to this problem by pointing to the fact that at some point we just need to stop asking 'why' questions. Compare: we found out that water, i.e. the stuff that is

liquid at room temperature, that quenches thirst and that is ideal for washing, in fact has the chemical structure H20. It could have had a different structure, but as it turns out, it doesn't. Identity theorists point to the 'fact' that it doesn't make sense to ask *why* water is $H_2O$ rather than some other chemical structure. It is just something we found out to be the case. And so, they argue, it is with mind-brain identification.

The problem with this comparison, however, is that the identification of water with $H_2O$ is not as inexplicable as it may at first seem. Water can be defined by all its properties: it freezes at 0 degrees Celsius, it boils at 100 degrees Celsius, it is transparent and liquid at room temperature, sugar dissolves in it, it cannot be mixed with oil, etc. etc. All these properties can be explained by reference to the properties of $H_2O$ molecules. Hence we explain *why* water is $H_2O$. A similar explanation is lacking in the case of brute mind-brain identification. We may have discovered that mental states are brain states, but this still leaves room for the question of why that is the case.

## 1.4 Functionalism

The identity theory was boasted of as the theory of the future – in the middle of the previous century. In fact much of the present-day neuroscientific literature reads as if the theory is naturally presupposed. However, in philosophy the theory was superseded by its successor, functionalism, within fifteen years. The main reason for this is that functionalism offers clear solutions to the two main problems that plague the identity theory: 1. How is multiple realisation possible? And, 2. How can we explain why some brain states are mental states? The key to these solutions is to adopt a strategy that was borrowed from Ryle's behaviourism: start by defining what we mean by the various mental-state concepts we have. If we want to know how a belief state can be a brain state, we need to start with being clear about what we mean by 'belief'. But while adopting this fruitful strategy of first defining what we want to explain from behaviourism, it goes beyond behaviourism

in being able to acknowledge, explicitly, that there is an inner aspect to the mind.

The central idea behind functionalism is that we should not characterise mental states in terms of what they *are* (either made of soul-stuff or made of brain-stuff). Rather, we should characterize them in terms of what they *do*. A mental state, according to functionalism, is characterized by how it is caused (e.g. by which sensory inputs) and what it causes (e.g. what behaviour or what further inner changes). This move has proven to be remarkably powerful.

### *1.4.1 behavioural dispositions as brain states*

One of the first steps towards functionalism was David Armstrong's (1926-2014) idea that logical behaviourism and identity theory need not necessarily be regarded as opposing theories. Behaviourists such as Ryle define our mental states in terms of our dispositions to behave in specific ways in specific circumstances (see 1.2.3). Behaviourists talk about such dispositions as if they are abstract properties of persons, that is, properties that do not exist, concretely, at some specific physical time and place. According to this way of thinking, dispositions are 'if-then' properties that cannot exist as concrete physical parts of the body. They exist as properties of the person as a whole.

Armstrong convincingly shows that this view of the nature of behavioural dispositions is wrong. Compare the fragility of glass. Fragility is the disposition to break under specific circumstances. That disposition is physically realized in the micro-physical structure of glass. It is the specific distribution of the molecules glass is made up from that determines that when sufficient force is exerted glass will break. In exactly the same way, Armstrong claims, our tendencies to behave in specific ways under specific circumstances are determined by our brain states, for the brain determines our behaviour dependent on the sensory inputs it gets about the circumstance we are in. If the mind consists of behavioural dispositions, then, the mind resides in the brain.

The identification of behavioural dispositions and brain states is a first step towards a solution of one of the main problems facing

logical behaviourism. Behaviourism appears unable to account for the idea that there is an inner aspect to the mind that is hidden behind our behaviour. But if behavioural dispositions are, indeed, brain states, then they can really exist internally to the person, not expressed in overt behaviour. An actress who feigns pain behaviour (see 1.2.4) displays that behaviour on the basis of different brain states than the person who really is in pain. And the person who is able to withstand showing her pain in overt behaviour is nonetheless in a brain state that would, in the absence of inhibition, lead to normal pain behaviour.

### *1.4.2 Folk-psychology and mental holism*

The fact that behaviourism cannot handle the idea that there is an inner aspect to the mind is not the only problem faced by that theory. As we saw (in 1.2.4) mental holism poses a serious problem as well: specific behaviour in a specific situation allows for many competing psychological explanations in terms of different sets of mental states. Wilfrid Sellars (1912-1989) was the first to recognize this problem. He agreed with Ryle, however, that the main purpose of ascribing mental states to people is to characterize and or predict their behaviour. His proposal, therefore, was to abandon the idea that there is a one-to-one definitional relation between mental states and behaviour and instead allow for a more hypothetical explanatory relation between mental states and behaviour. Thus he introduced the idea of what is now known as 'folk-psychology', which formed a second impulse for functionalism.

As we will see in Chapter 8, the term 'folk-psychology' is currently widely used in the philosophy of social cognition, sometimes in a fairly loose way. For Sellars, however, the term refers to our system of psychological concepts that is used in our everyday practice of ascribing mental states to each other and ourselves. Folk-psychology in this sense is a folk-theory. Ascribing mental states to people, Sellars argues, has a function that is comparable to postulating theoretical entities in science. Suppose, for example, that a scientist observes a new planet. The orbit of this planet deviates from what her knowledge about planets so far suggests. A

good option in such a case is to postulate the presence of a large mass, such as e.g. a black hole, that pulls the new planet out of its 'normal' orbit. The scientist cannot observe this hypothetical mass, but is assuming its presence explains the observed trajectory of the new planet. The postulated mass is a theoretical entity; its existence is a theoretical hypothesis. Sellars' claim is that the ascription of mental states is very similar, it is the postulation of unobserved entities in order to explain and predict observed and observable behaviour.

The meaning of our mentalistic language, then, can be explained in terms of the predictive and explanatory function of folk-psychology, in Sellars' view. In that sense, Sellars may be regarded as a behaviourist. But his kind of behaviourism does not suffer from the holism problem: the idea of folk-psychology leaves room for various explanations of the same behaviour – just as it is open to the scientist in the above example to consider alternative explanations for the deviant orbit of the newly observed planet. Accepting mental holism, and hence the possibility that one instance of behaviour can be explained in terms of many different sets of mental states, need not necessarily mean that we can never really know what mental states people are in, for every ascription of sets of mental states yields predictions. Further behaviour of the person at issue will thus diminish the range of possibilities. In practice there usually aren't that many rival psychological interpretations of somebody's behaviour.

The idea of folk-psychology makes a little more room for the idea that the mind is an inner realm than does Ryle's behaviourism. For one thing, in Sellars' view mental states are not reducible to behaviour. More importantly, Sellars argues that we can and do apply the ability to wield folk-psychology to ourselves. This means that we learn to express our own behavioural tendencies in terms of what we call our thoughts, intentions, wishes, beliefs, etc. Thus we come to speak of our own inner lives. It is crucial here to note that this inner life is not the inner life that Descartes thought we had. In Descartes' view we directly introspect what goes on in our minds. Sellars leaves room for the fact that we may *think* this is

how we know our inner lives. This is because the first-personal use of folk-psychology has become second nature to us. But even so, it remains a theoretical interpretation of our selves and our behaviour, rather than introspective access to a real inner realm.

### *1.4.3 The basic idea: causal roles and neural realizations*

Keeping in mind Sellars' idea of folk-psychology as a folk-theory, David Lewis (1941-2001) formulates a classic version of what is later to be called 'analytical functionalism' in two papers, published in 1966 and 1972. In Lewis' account the meaning of mental-state terms – 'belief', 'desire', 'will', 'thought', 'feeling', etc. – depends on the ways in which these terms are interconnected into a framework, a theory, that explains, predicts, describes and makes sense of our behaviour in everyday life. The theory differentiates between different mental states in terms of the way in which a state is caused by sensory or other inputs, and the internal and behavioural effects it has. Intentions, for instance, tend to be produced by a combination of desires and beliefs, and they produce actions. Beliefs, in turn, tend to be produced by sensory or verbal inputs, and in combination with other states they may form further beliefs but also hopes or fears as well as intentions and actions. We can give similar causal profiles for all other kinds of mental state. In short, as it is put by Lewis, each mental state is characterized by the *causal role* it plays within the person whose state it is.

Of course the causal-role description of intentions and beliefs we gave above is laughably course-grained and imprecise. For present purposes it suffices, though. It is instructive to point towards the three elements of which most causal roles that characterise mental states are built. An example may help. Take a state of being in pain. Normally such a state has a specific cause such as damage to bodily tissue (that causes so-called 'nociceptors' to fire). It also produces characteristic behaviour such as wincing, screaming, grabbing the sore spot or calling a doctor or seeking a band-aid. It produces specific internal effects such as tending to make much other mental activity impossible. Most mental states can be characterised in

terms of these three elements: 1. Inputs, i.e. what causes them; 2. Outputs, i.e. the behaviour they cause; and 3. Consequences for the internal state of the person, which in turn can be understood in terms of this person's changing susceptibility to external (and internal) influences; (for instance, you will miss the subtleties of the Bach Cantata you were listening to after boiling-hot coffee is accidentally spilled over your arm).

The idea of mental states as causal-role states is quite revolutionary. It allows us to define mental states in terms of what they *do*. Compare the concept of a carburettor. Carburettors are parts of the engines of cars that are also defined in terms of what they do: they mix fuel with air so that it becomes combustible. This distinguishes carburettors from other parts of the engine such as the cylinders where the actual combustion takes place or the pistons that are pushed outwards by the combustion. Just as carburettors are distinguished from other parts of the engine by the causal role they play, so beliefs are distinguished from desires, hopes and other mental states by the causal role they play. Mental states can be defined in terms of their – interconnected – causal roles.

While Sellars considers mental states as theoretical hypotheses, Lewis connects more directly to the identity theory and Armstrong's idea that behavioural dispositions are in fact realized in our brains. This means that the causal roles that define our mental states are thought to be played by concrete brain states. Put differently: causal-role states – mental states – are *realized* by brain states. When your pain is caused by spilling hot coffee, while it causes you to scream and not pay attention any more to the Bach Cantata you were listening to, there is actually a specific brain state that is caused by the damaged skin tissue (via nociceptors and other nerves), that causes you to scream and to not be able to listen to Bach any more. That brain-state *is* your pain because it realizes (or plays) the causal role that defines your pain.

The subtle, but as it will turn out, crucial, difference between this view and the standard identity theory is that my brain state is my pain state *only* in virtue of its functional properties, i.e. the role it plays in me, hence the name 'functionalism.'

### *1.4.4 Why mental states can be brain states and vice versa*

Functionalism solves both problems that confront classical identity theories. Let's start with the fact that according to classical identity theories we need to accept that mental states are brain states as a brute, unexplainable fact of nature. Functionalists have a better story. By defining mental states in terms of the causal roles they play and by subsequently explaining how our brain states play the causal roles that are definitive of mental states, we can explain why certain brain states are mental states: they play the right causal role.

The relation between a mental state and a brain state, according to functionalism, is not mere unexplainable brute identity. Rather, it is a relation between a causal role and its realizer. In order to grasp the idea behind role-realizer distinction, a comparison may help. When Gregor Mendel postulated the laws of genetics, at the end of the nineteenth century, he postulated theoretical entities that are the bearers of hereditary traits. He called these unknown entities 'genes'. For a long time no one knew what genes were. They only knew what genes *did*. Genes are defined in terms of their causal roles (and these causal roles are extensively described by the laws of genetics – just like folk-psychology describes the causal roles of our mental states). It was only in the early 1950's that Francis Crick and James Watson discovered the chemical structure of genes. Genes turned out to be DNA, or, better put, DNA turned out to be the realizer of the gene-roles.

By defining the relation between mental states and brain states in the same way as the relation between genes and DNA, functionalists are able to explain why mental states are brain states and *vice versa*.

### *1.4.5 Multiple realisation explained*

The second problem that faces the classic identity theory is multiple realisation. How can one type of mental state be realised by different types of brain state? How, for instance, can different types of animals all be in pain when they do not have a shared neural make-up? Functionalism is able to explain this by means of the role-realiser distinction. The same causal role can be played

by different realisers. Suppose we define pain in terms of 'being caused by tissue damage, causing screaming and the causing of decreased susceptibility to other sensory inputs'. This causal role can be played by many different types of brain state in many different types of brain. Functionalism allows us not only to accept multiple realisation, like the token-identity theory, but also to explain it.

## 1.5 Mind as a computer program

The idea that machines might be able to think is an old one – goes back at least to the beginning of the nineteenth century when Lady Ada Lovelace (1815-1852) and Charles Babbage (1791-1871) developed what they called 'the analytic engine.' The first scientist to propose the idea of a universal programmable and – programmable machine, however, was the British mathematician Alan Turing (1912-1954). Is it really possible that such a machine – a computer – can actually think? In 1950 Turing published a paper in which he explains why he thinks the answer to this question is 'yes.' That paper spawned the research program we now know as artificial intelligence or AI.

Apart from mathematical ideas, philosophical theorizing plays a leading role in this movement. Turing's idea of a computer program – so familiar to us now that it is hard to imagine how revolutionary it once was – is central to the idea of a thinking machine. But can we really define 'thinking' in terms of a program? Functionalism allows us to answer that question affirmatively for it defines mental states in terms of the functional properties of mental states rather than the mere physical properties of the realizers of causal – functional – roles. Thus 'thinking' can be defined in abstraction from body, brains, making it possible to conceive of other systems as being capable of thought as well. Should we be able to build a functional equivalent of our brains in a computer, then functionalists would have to admit that this computer can think just as well as our brains can.

However, the development of more and more intelligent computers allowed philosophers to reverse the comparison. According