1 Introduction

It seems obvious that we humans have mental states such as thoughts, beliefs, desires, sensations, and emotions – or at least as obvious as that we have physical states such as height, weight, and hair color. It also seems obvious that changes in our physical states can cause changes in our mental states, and vice versa: touching a hot stove will (normally) cause us to feel pain, think that we've been careless, and want the pain to stop – which, in turn, will cause us to wince, or curse, or run to the freezer to get some ice. However, although these claims may seem obvious, they raise a number of questions that do not have obvious answers.

Some of these questions about mental states are *metaphysical* questions: What kinds of things are thoughts, desires, sensations, and emotions; what is their relation to the physical states of our bodies and brains; and how do changes in physical states produce changes in mental states, and vice versa? Another metaphysical question is whether nonhuman creatures can also have mental states, and if so, which creatures – for example, chimps, cats, octopuses, androids – and which types of states – for example, sensations, thoughts, hopes, fears?

There are also *epistemological* questions in this domain; among them are how we could know whether chimps, cats, octopuses, androids – and indeed humans other than ourselves – have mental states, and if so, whether those states are similar to our own. Indeed, there are questions about how we come to know about our *own* mental states and whether we could ever think that we think something, want something, or feel something – and be wrong. There are *moral* questions in this domain as well; among them are whether creatures that can think or feel should be treated differently from those that cannot and whether we are entitled to expect creatures that can think or feel to treat us in certain ways.

This Element focuses on the metaphysical questions. However, as will become clear, the answers to the metaphysical questions have implications for the others – and vice versa – and so these other questions cannot be ignored entirely in evaluating different theories about the nature (and extent) of mental states and their relation to the physical world. Thus, while these questions will not be the focus of attention, they cannot be completely ignored.

On the other hand, this Element does not aim to present every metaphysical theory of mind that has been proposed throughout the history of philosophy but focuses instead on five that – at least arguably – are taken most seriously in contemporary work on the subject. These are *Dualism*, *the (Mental–Physical) Type Identity Theory*, *Role Functionalism*, *Russellian Monism*, and *Eliminativism* (or Illusionism).

Dualism is the thesis that the mental and the physical are two distinct kinds of things, each independent of, and irreducible to, the other. Dualism has three major varieties: *Substance* (or *Cartesian*) *Dualism*, the thesis that physical and mental states are states of two distinct types of substances such as material substances that occupy space and are capable of motion and immaterial substances that exist outside of space; *Property Dualism*, the thesis that while there are no immaterial substances, there are certain sufficiently complex material substances (such as human brains) that have states or properties that are irreducibly mental; and *Panpsychism*, the thesis that all things, from atoms to humans to planets, have both physical and irreducibly mental states.

The *Type Identity Theory* is a species of *Physicalism*, the thesis that minds and mental states are nothing over and above bodies and physical states. Evaluations of Physicalism often include debates about whether there is *anything* in the world that is nonphysical, for example, immaterial gods, ghosts, or even numbers. The concern here is narrower, however, and focuses on whether (human) mental states, and all their properties, are physical. The Type Identity Theory is the claim that each type of mental state (e.g. the feeling of pain, the thought that today is Friday, the desire for chocolate) is identical with some type of physical state, presumably some state of the brain and central nervous system, for example, that pain is identical with a certain sort of C-fiber stimulation. Type Identity statements, therefore, can be true only if all (and only) instances of a particular type of mental state (e.g. pain) are instances of the same physical type (e.g. C-fiber stimulation).

There is another species of Physicalism, *Nonreductive Physicalism*, that does not require type identity, but only that each particular instance (or "token") of a type of mental state be identical with a token of some type of physical state or other. On this view, creatures with physical states very different from our own could nonetheless have the same mental states as we do, as long as we have certain other properties in common that are not irreducibly mental. Different Nonreductive Physicalist theories specify different properties to play this role, but the most common species of Nonreductive Physicalism is *Role Functionalism*.

Role Functionalism is the thesis that what makes something a mental state is not its internal constitution, but the role it plays, the function it has, in an individual's psychology. Role Functionalism too has a number of varieties. These arise from differences in which sorts of roles are viewed as definitive of mental states and which sources of information can be used to specify those roles. Common Sense (or Analytic) Functionalism requires that information be derived from our commonly held "platitudes" about the causal roles of mental states in the production of other mental states and behavior, while

Psychofunctionalism (or Empirical Functionalism) permits information from empirical psychology and neurophysiology to have a role in characterizing mental states, even if it is not commonly known.

All versions of Physicalism, in contrast to Dualism, are species of Monism, the thesis that there is just one fundamental type of thing in the world from which everything else is derived. There are other species of Monism, including *Idealism*, the thesis that the fundamental constituents of the world are minds and their perceptions, thoughts, and volitions, and *Neutral Monism*, the thesis that the fundamental constituents of the world are neither physical nor mental, but rather "neutral" properties from which both physical and mental states arise. Although there are different varieties of Neutral Monism, one of the most interesting for contemporary philosophers is Russellian Monism.

Russellian Monism derives from Bertrand Russell's (1927) view that the physical sciences describe only the structural or dispositional properties of the things that occur in nature and that these dispositions must be grounded in, or underlain by, intrinsic or categorical properties. In Russell's view, these categorical properties not only ground the dispositional states described by the physical sciences but also provide the basis of our conscious experiences.

Although Dualism and the various species of Monism may seem to exhaust the possibilities for a theory of minds and mental states, there is one further theory that warrants discussion: Eliminativism.

Eliminativism (or Illusionism) is the thesis that there are no such things as mental states and properties – or at least no states that possess certain essential features that we commonly assume sensations, perceptions, beliefs, or desires to have. Some Eliminativists direct their skepticism to sensations and perceptual experiences; others to beliefs and desires – and they do so for importantly different reasons.

This may seem to be a tidy categorization of the available theories of mind, but there is some debate about which views belong to which categories; is Nonreductive Physicalism really a species of Physicalism; is Neutral Monism genuinely neutral; is the Type Identity Theory just Eliminativism in disguise? Even more broadly, there is debate about the proper characterization of Physicalism itself.[1]

Moreover, one need not expect any one theory to provide the best account of all types of mental states; one can pick and choose among them, giving up unity for plausibility. For example, one can, and many do, endorse Role

[1] Should physical states be characterized as the states described by our *current* physical (and chemical and biological theories) or by the theories that will emerge at the end of scientific inquiry? See Hempel, 1969. This debate, however, does not affect this discussion of the metaphysics of mind.

Functionalism as a theory of thoughts, beliefs, and desires but prefer the Identity Thesis – or even Dualism – as an account of perceptual experiences and bodily sensations. Or one can adopt Eliminativism for beliefs and desires but endorse the Identity Theory (or Dualism or Functionalism) for sensations and perceptual experiences.

In the sections to follow, I will present these five theories of the nature of mental states and sketch their primary strengths and weaknesses. In doing so, I will pay special attention to how well they account for what seem to be the distinctive properties of states such as sensations and perceptual experiences, namely, their *qualitative (or phenomenal) character*, or, in Thomas Nagel's (1974) now-classic locution, *what it is like* for someone to be in those states. I will also focus on how well these theories capture what seem to be the distinctive properties of states such as thoughts, beliefs, and desires, namely, their capacity to *represent* – or *be about* – items in the world.

Here too, however, there is no tidy categorization; many philosophers argue that even though there are important differences between thinking and feeling, sensations and perceptual experiences can represent items in the world (or in one's body) in addition to having qualitative character, and some argue that thoughts can have specific qualitative characters in addition to representing items in the world. Indeed, some argue that sensations *must* be representational and thoughts *must* have a qualitative character.[2] These views are contentious, but if they are true, then the problems raised for sensations and perceptual experiences will extend to thoughts – and vice versa.

As will (or should) become clear, there is no knockdown argument for or against any of these theories, and they all have features, or implications, that may violate our commonsense, pre-theoretical views about what mental states are, what sorts of creatures can have them, and what their relation is to bodily states. This has prompted some (e.g. Schwitzgebel, 2014) to question whether it's rational to accept any of these theories – even while recognizing that one (or some combination) of them has to be true, since they exhaust the possibilities. In my estimation, this verdict is too pessimistic. Although I will try to give a fair account of the strengths and weaknesses of all the views in question, readers may weigh these strengths and weaknesses differently and judge that there are good (if not airtight) reasons to accept one or the other of these views.

[2] For a good example of the first, see Byrne (2001), and of the second, see Horgan and Tienson (2002).

However, in the final section, I will (briefly) present some further, recently articulated, questions about the relation between mental states and individual brains and bodies that arise for almost any theory of the metaphysics of mental states. These questions are puzzling and important, and they are just beginning to be discussed. Thus, even for those who clearly prefer one (or some combination) of the theories discussed in these sections, there is still a lot of work to be done to determine the relation of mental to physical states and the place of the mind in the natural world.

2 Dualism

There are many varieties of Dualism – the thesis that mental and physical states are distinct from and irreducible to one another – but all hold that in a world containing nothing but physical objects, events, and properties, there would be no creatures with thoughts, sensations, volitions, or any other sort of mental states. For that, the world must include something more.

But what is this "something more"; which creatures possess it; and what is the relation between whatever it is and the world of physical objects? These are questions that the different varieties of Dualism answer in different ways.

According to Substance Dualism, for there to be creatures with mental, as well as physical, states and processes, the world must include immaterial substances – minds (or equivalently, souls) that can think, perceive, and will – in addition to the bodies that take up space, have the capacity to move, and can be perceived by the senses. Although this was the dominant view in the ancient and medieval world, in contemporary discussions, it is associated primarily with Rene Descartes (and often called Cartesian Dualism) – for two reasons. One is that Descartes was among the first to characterize the mind as we now conceive it, namely, as the locus of conscious mental activities exclusively (i.e. thinking, feeling, and willing), rather than those activities plus others distinctive to living things, such as locomotion and respiration. Another is that Descartes's most influential argument for the distinction between mind and body, which he presents in Meditation Six of his *Meditations on First Philosophy* (1641/1984), provides the template for the most influential contemporary arguments for Dualism, and the responses to this argument by Descartes's own interlocutors provide the template for the most influential responses to those contemporary arguments.

This argument, in brief, is:

1. I can clearly and distinctly understand myself to exist apart from my body.
2. If I can clearly and distinctly understand *x* to exist apart from *y*, then it is possible for *x* to exist apart from *y*.

3. If it is possible for *x* to exist apart from *y*, then *x* is not the same thing as *y*.

THEREFORE
I am not the same thing as my body.[3]

Many find the conclusion of this argument attractive since it opens up the possibility that one's mind, and thus one's self, could be immortal – or at least continue to exist for some time after the destruction of one's body. The premises of this argument, however, need explication – and defense.

For Descartes, to have a clear and distinct understanding of something is to have a conception of it that reveals its nature or essence, that is, the properties it must possess in order to exist – and we attain clear and distinct understanding by considering our ordinary idea of something and thinking carefully about which properties it can lose and which it must retain to remain that very same thing. Descartes defends premise 1 by arguing that he has a clear and distinct conception of himself as (essentially) something that thinks (i.e. as the locus of conscious mental activity) and a clear and distinct conception of his body as (essentially) something that occupies space – and that it is perfectly coherent to think of things that occupy space as lacking conscious mental activity, and vice versa.[4]

This premise may seem plausible, at least initially. But some of Descartes's contemporaries, most famously Pierre Gassendi, argued that Descartes's reports of his own clear and distinct conceptions may just be wrong. As Gassendi puts it in his Objections to *Descartes's Meditations* (1641/1984), Descartes's supposed understanding of himself as an immaterial substance may really be a conception of himself as "a wind, or rather a very thin vapour . . . diffused through the parts of the body and giving them life."[5] Gassendi's question, in short, is whether we can be wrong about what the contents of our clear and distinct conceptions *are* – and as will become clear, this worry informs contemporary discussions of the metaphysics of mind as well.

[3] This argument is extracted from the following passage:

> First, I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it. Hence the fact that I can clearly and distinctly understand one thing apart from the other is enough to make me certain that the two things are distinct, since they are capable of being separated, at least by God . . . [Now] on the one hand I have a clear and distinct idea of myself, insofar as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of body, insofar as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it. Descartes (1641/1984, p. 53).

[4] However, there are versions of Substance Dualism in which immaterial substances can have spatial location. See Hart (1988) and Latham (2001).

[5] See Gassendi (1640/1984), p. 180. The argument gets quite heated – on both sides: see Descartes's Reply to Fifth Set of Objections 1641/1984): 241-267.

Premise 2 may also seem dubious, especially since Descartes defends it by appealing to the existence and nondeceptiveness of God (which he claims to have proven in, respectively, in his *Third* and *Fourth Meditation*). If God exists and is truly nondeceptive, Descartes argues, then we must possess some sort of faculty that, if used correctly, will get us to the truth. And what better candidate could there be for such a faculty than clear and distinct understanding!

Here too, however, Descartes's contemporaries, most famously Antoine Arnauld, raised questions about the connection between clear and distinct understanding and possibility. In his Objections to *Descartes's Meditations* (1641/1984), Arnauld argues that there is an obvious counterexample to premise 2, namely, that someone can clearly and distinctly understand that a triangle is right-angled, without needing to understand that it obeys the Pythagorean theorem, and thus, it would follow that there can be "a right-angled triangle with the square on its hypotenuse not equal to (the sum of) the squares on the other sides" (1641/1984, p. 182) – which is obviously impossible.

Descartes's response to this objection is to argue that to have a clear and distinct understanding of something, one needs to have a sufficiently "complete" conception of it to guarantee its existence. And when one has a sufficiently complete conception of what it is to be a right triangle *and* what it is to be a figure that obeys the Pythagorean theorem, one will be able, at least in principle, to see that if one exists, so must the other – and so this case is no counterexample to premise 2.

However, even if Descartes's response to Arnauld is convincing, contemporary thinkers may be reluctant to embrace a principle that makes the link between what we can conceive and what is possible dependent on the existence and nondeceptiveness of God. But even if they are skeptical of Descartes's defense of this premise, many contemporary thinkers agree that if there is any way we understand the world that can provide reliable evidence for claims about what is possible – and thus about the nature or essence of things – it will be something like Descartes's clear and distinct understanding. In achieving this sort of understanding, we've tried our best and thereby have the best possible evidence for such claims!

So if Substance (Cartesian) Dualism were true, it would support the firm and widespread intuition that the distinctive features of our conscious mental states are so radically different from any physical states, events, or properties – either macroscopic or microscopic – that they could not be (or be explained by) anything exclusively physical. This has come to be known as the "hard problem of consciousness" (Chalmers, 1995). There are many examples of this intuition

throughout the history of Western philosophy, for example, Leibniz's argument that no physical substance whose workings can be explained by mechanistic principles could possibly think or perceive:

> If we imagine that there is a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters a mill. Assuming that, when inspecting its interior, we will find only parts that push one another, and we will never find anything to explain a perception
>
> (1714/1991, section 17).

Another well-known example is T.H. Huxley's skepticism about the possibility of a neurophysiological explanation of conscious experience:

> How it is that anything as remarkable as a state of consciousness comes about as a result of irritating nerve tissue, is just as unaccountable as the appearance of the Djin, where Aladdin rubbed his lamp in the story (1875).

There are many other examples in between – and since – among them is McGinn, who asks: "How is it possible for mental states to depend upon brain states? How can technicolour experience arise from soggy grey matter?" (1989, p. 349).

However, although Substance Dualism has its attractions, it also has serious problems. First, it introduces a new type of substance, immaterial minds, into the world and thereby raises questions about when these substances get created, whether (and if so, how and why) they may be destroyed, and how many (and which sorts) of one's mental states (ideas) they may retain after the death of the body. More generally, one may wonder whether there are more parsimonious ways to explain how humans think, feel, and act; can this be done without appeal to immaterial minds? Moreover, if minds and bodies are distinct substances, it is hard to explain the seeming unity of mind and body that we experience in ordinary human life when our sensations and beliefs seem immediately and inextricably linked to what is going on in our bodies. Descartes himself recognizes the difficulty, later in Meditation Six of his *Meditations on First Philosophy* (1641/1984), where he acknowledges that:

> Nature . . . teaches me, by these sensations of pain, hunger, thirst, and so on, that I am not merely present in my body as a sailor is present in a ship, but that I am very closely joined and, as it were, intermingled with it, so that I and the body form a unit. If this were not so, I, who am nothing but a thinking thing, would not feel pain when the body was hurt, but would perceive the damage purely by the intellect, just as a sailor perceives by sight if anything in his ship is broken.

This intermingling is beneficial, Descartes suggests, because having sensations of pain and pleasure is an effective way of getting information about the helpful and harmful conditions in our environment. But it is hard to take these remarks about intermingling literally, since it is hard to understand how an immaterial substance that does not occupy space could "intermingle" with a body that does. Therefore, most theorists understand Descartes to be claiming that what unifies the mind and body of an individual is the existence of a unique, direct *causal connection* between that individual's physical and mental states: I put my hand near the burner on the stove, which causes me to feel heat and believe that moving closer would be painful, which, in turn, causes me to move my hand away. Such directness and (relative) immediacy occurs when, and only when, *my* bodily states produce *my* mental states (and vice versa) – and this is enough to explain the seeming unity of my mind with my body that I experience when I interact with the world.[6]

However, it is also difficult for a Substance Dualist to explain just how mind–body causation works. In *The Passions of the Soul*, Descartes proposes that "There is a little gland in the brain where the soul exercises its functions more particularly than in the other parts of the body" (1649/1984, section 31); that is, the pineal gland, which, as he puts it (1649/1984, section 32), serves as a kind of "funnel" by which neural activity in one's brain can have effects on the mind (soul) that correspond with those bodily (neural) activities. Causation also occurs in the other direction: the mind (soul) can act, via the pineal gland, to produce changes in one's brain that in turn have various differential effects on the body, such as moving its limbs.

Descartes's proposal, however, does not provide much insight into *how* this happens; how it is that neural activity can produce changes in an immaterial substance; and – even more puzzling – how an immaterial substance can produce changes in a body, especially given phenomena such as the conservation of energy and the principle that every physical event has a (sufficient) physical cause. This worry was expressed particularly forcefully by one of Descartes's contemporaries, Princess Elisabeth of Bohemia, who in a long correspondence with Descartes (1643–1647) asks how "the soul of a human being [which] is only a *thinking* substance . . . can . . . affect the bodily spirits, in order to bring about voluntary actions," given that causation requires "*contact* between the two things, [which in turn] requires that the causally active thing be extended" (1643/2017, p. 1).

[6] Some scholars, however, argue that Descartes regards humans as a third kind of substance and that they should be regarded as "trialists" rather than dualists. Among others, see Hoffman (1986).

Descartes (1643/2017, p. 3) responds by suggesting that the problem arises because we have confused the notion of "the soul's power to act on bodies with the body's power to act on other bodies" – and that soul–body causation is a different sort of process that requires neither contact nor energy transmission. However, he never says just what this process is, but only that something like this *has* to occur, since he's proven that minds (souls) are distinct from bodies, and we know that there is causal interaction between them!

Clearly, one way to resolve this dilemma without abandoning Dualism is to give up the claim that there is causal interaction between mental and physical states. Some of Descartes's (near) contemporaries do just this and argue that although it may seem that the mind has an effect on the body (and vice versa), this is an illusion. One view, Parallelism, contends that mental and physical events occur in perfect parallel and do not interact at all – due to God's having set things up this way and then letting things in each domain unfold through time in preestablished harmony. Another view, Occasionalism (associated primarily with Malebranche), contends that when it seems as though a physical event in my body is causing something to go on in my mind, what is *really* happening is that God is taking the occurrence of that physical event to be an occasion for producing some mental state in me.

These alternatives will not be pursued further, since their problems are no doubt salient. But it is important to recognize that at least some Dualists at the time were willing, in one way or another, to abandon the claim that there is causal interaction between mental and physical states. As will become clear, the worries expressed in this seventeenth-century dialectic, especially about how mental states (as understood by Descartes) could cause bodily states, remain major worries for Dualism even today.

A further question for Substance Dualism concerns just which physical creatures possess immaterial minds. Descartes (1637/1984) himself argues that humans are the only creatures with minds as well as bodies, on the grounds that – as he argues in the *Discourse on Method* (and other places) – humans are the only (mortal) creatures capable of thought and rationality. We know this, he argues, because only humans can use language and behave appropriately in a wide range of situations, and it is these capacities that, as Descartes puts it, distinguish "man from beast," and also human from mere machine. Admittedly, there has long been a dispute about whether all nonhuman animals lack the linguistic capacities and response flexibility that Descartes requires for having a mind, and there is increasing debate about whether machines could someday possess them. But it is hard, nonetheless, to settle on criteria for what it takes to make the cut.