0. Philosophy of Mind and the Mind-Body Problem

Philosophy of mind and the mark of the mental

The philosophy of mind is an area of philosophy delineated by a set of problems or philosophical questions about the mind, and mental phenomena. Even to say this much, however, is to say something that is not altogether without controversy. There is, after all, already a question to be asked here about how we should characterise things (events, actions, states, properties, processes, phenomena) as being mental. And this is a question that has not been answered by everyone in the same way.

An obvious starting place, perhaps, is to roughly define it as the stuff that goes on in our heads – things like mental arithmetic calculations, daydreaming, planning revenge, or thinking about yesterday’s roast dinner. But this doesn’t seem quite right. After all, it seems like there are lots of other phenomena that we would probably want to class as mental that don’t seem to fit this description. A feeling of pain when I knock my shin, for instance, or my growing feeling of hunger, or a ticklish feeling between my shoulder blades that I can’t reach.

One thing that does seem to be in common to all of these things is that there is something it is like to experience them, so perhaps this is the way to pick out the mental things. This is more promising, but we still might not think that this is going to be enough as our full answer – there’s nothing it’s really like to experience the belief that Paris is the capital of France, but surely beliefs are exactly the kind of thing we think of as mental.

Another candidate might be something that philosophers call intentionality. A state is intentional when it is of or about something. My belief, for instance, is about Paris, and my daydream might be of moving to New York. But what about my hunger feeling? Is that really of or about something?

There is obviously much more to be said about what we should count as the mark of the mental, and brainstorming this question might be a nice place to start students thinking about the philosophy of mind. In what follows, however, I won’t further pursue this question any further. I trust that we have a rough-and-ready understanding of what minds and mental phenomena are that will do for our purposes. Moving away from this question, the focus of these notes will be one of the central problems in the philosophy of mind. This is the problem called the mind-body problem.

The mind-body problem

Questions in the philosophy of mind generally fall into two categories. The first is a category of epistemological questions that ask about what we can know when it comes to the mind. These include questions about how we can know about the mental states of others, how we can know about our own mental states, and the nature of these forms of knowledge. The other category consists in metaphysical questions that ask about the fundamental structure of the mind. The problem that has dominated the philosophy of mind over the last three centuries or so, and which will provide the basis of these notes, is a metaphysical one: the mind-body problem.
On the face of it, the mind-body problem seems to be a fairly simple one. It amounts to the challenge to say what the relation is between our minds, or mental states, and our bodies, or brains. Why should we be moved to consider this problem at all?

The real force of the problem comes from two sets of intuitions, both of which seem pretty compelling. The first of these is the intuition that the mind and the body are two very different kinds of things. The body is an extended object in the physical world – it has a size, a weight, a shape, a spatio-temporal location, I can feel it, taste it, smell and see it. And as a physical object in the world, it is just as well placed as other worldly physical objects to causally interact with the things around it. My body gets onto trains, sits on chairs, picks up coffee cups, and shakes hands with other bodies. Moreover, all of this is done in a way that would seem to be perfectly respectable to a physicist. Just like other objects in the world, our bodies conform to the natural and physical laws – the laws of gravity, causation, motion and so on. In short, our bodies just seem to be physical things like any other. And the same, I take it, can be said of our brains.

The mind, by contrast, does not seem to be like this at all. Our mental states do not seem to be the kinds of things we could locate in the physical world or encounter through the senses – to talk in this way just seems to get the sort of thing we’re dealing with wrong. If we start asking whether our mental states are subject to laws of gravity, or could get onto trains, or weighs more than a bag of sugar then it just looks like we’ve made a mistake. As pointed out above, it’s not east to say just what properties we can truly ascribe to our mental states, but it will probably be something like consciousness, intentionality, subjectivity, or the fact that there’s something it’s like to be in that state – all features that seem to have no place in our purely physical description of our bodies.

At first sight, then, it looks like our bodies and our minds are very different kinds of entities. It is, on the other hand, plain to anyone who has ever had both a mind and a body that they are more deeply and inextricably bound up together than this intuition of distinctness might at first lead us to believe. My body isn’t just a vehicle for my mental life. Every thought I think, every action I perform, or sensation I experience, rather, involves a complex and indissoluble engagement of both mind and body; as the founder of contemporary philosophy of mind, René Descartes, famously put it: ‘I am not only lodged in my body as a pilot in a vessel, but that I am besides so intimately conjoined, and as it were intermixed with it, that my mind and body compose a certain unity’. If we take the intuitions of the last two paragraphs too seriously, however, then we are in danger of slipping into just this kind of distinct picture of the mind-body relation.

What, then, is the right way to understand the metaphysical relation between my mind and my body that will accommodate both the intuition of distinctness and the thought that our minds and bodies are deeply interconnected? This is the mind-body problem.

**Four kinds of response to the mind-body problem**

In these notes I will outline four ways you might try to respond to the mind-body problem.
The first, **substance dualism**, says that the mind and the body really are distinct kinds of thing. The body is made up of a physical substance, whereas the mind is something more like a ghostly soul, and as such is made out of a kind of ghostly mental substance which is non-physical. In the first part of these notes I will outline substance dualism, and set out some arguments both in support of it and against it.

The second kind of response I will consider is called **reductive materialism**. According to reductive materialists there is only one kind of substance in the world: the physical kind. When we talk about mental states and events, on this view, we’re really talking about physical states, even if we don’t always realise it. In the second part I’ll outline the core commitments of a reductive materialist, and then critically present two such accounts: *logical behaviourism* and *identity theories*.

A third option is a kind of compromise between the first two. Supporters of **non-reductive materialism** agree with reductive materialists that there is only physical substance in the world. They disagree with the reductive materialist, however, in the claim that the physical stuff is *all there is* to mental phenomena. Even if our mental states and events depend on, or are made up out of, our physical states, facts about our mental lives can’t be reduced to facts about our physical make-up. In part three I will summarise the claims characteristic of non-reductive physicalism, then critically present two such views – *anomalous monism* and *functionalism*.

A final response to the mind-body problem is called **eliminative materialism**. Like the other two forms of materialism, eliminative materialists only believe in physical substance. Eliminativist materialists, however, are far more radical than any of the other forms of materialism. This is because according to that view, mental phenomena are not only dependent on, or even reducible to, the physical domain; according to the eliminativist, mental phenomena don’t even exist! I will critically present this view in the last part of these notes.

**Teaching tips**

- If you’re planning to teach a course on theories of the mind, it’s crucial to get the students to see the force of the mind-body problem, and be persuaded by both sets of intuitions so that they will understand why there is a puzzle here in the first place.

- That the mind (or the soul) is distinct from the is an idea that is commonly found across different religions and societies – drawing on these kinds of examples could be one way to urge the first intuition of the mind-body problem.

- With regards to the second intuition, try to press the thought that we’re not so alienated from our bodies as to think that they are just receptacles for our minds – surely our minds and our bodies causally interact in all sorts of ways. Use lots of examples of mind-body causation, and of bodily sensations which seem to be both mental and bodily.
Discussion questions

- What do you think are the central defining marks of the mental? Can you think of any counterexamples to each other’s suggestions?
- In what ways do you think that the mind and the body or brain are different sorts of thing? In what way do you think that they are intimately related?
- Before turning to what various philosophers have said in response to the mind-body problem, what do you intuitively think is the relation between your mind and your body, or your brain?

Further resources

- Any introductory guide to the philosophy of mind will usually have a section on the mind-body problem or at least a fairly substantial account of it in the book’s introduction, so resources are plentiful on this topic. A few worth a look at include E.J. Lowe’s *An Introduction to the Philosophy of Mind (Cambridge Introductions to Philosophy)*, Cambridge: CUP (2000), William Jaworski’s *Philosophy of Mind*, Oxford: OUP (2011), or Tim Crane’s *A Philosophical Introduction to Minds, Machines and Mental Representations* New York: Routledge (2003), but this is only to name a few among the dozens available to choose from.
- Another, somewhat lighter, resource that might help with ideas about how to introduce or to discuss the mind-body problem is an episode of radio 4’s *In Our Time* with A.C. Grayling, Sue James and Julian Baggini, chaired by Melvyn Bragg. This is a lively discussion on the mind-body problem that traces it back to Plato and Aristotle, but also covers the problem’s more recent history and its Cartesian formulation: [http://www.bbc.co.uk/programmes/p003k9b8](http://www.bbc.co.uk/programmes/p003k9b8)
1. Substance Dualism

Duality of substances

According to supporters of substance dualism – sometimes also known as Cartesian dualism after the view’s most well-known supporter, the 17th century French philosopher and mathematician René Descartes – there are two kinds of substance in the world: there is the physical stuff, and there is the non-physical, mental stuff.

What differentiates the two? Well, a substance is a kind of thing that has, or instantiates, properties. We could identify the substance of water, for instance, by pointing to the kinds of properties it has (of being composed of hydrogen and oxygen, for instance, or of being liquid, or drinkable, or coming out of the taps and so on). These kinds of properties give us a way to hook onto the differences between different sorts of substances. If, for instance, we wanted to contrast the water substance-kind with another – say, the substance of gold – we could easily do so by appealing to their different properties. The substance of water has the properties listed above, while gold substance has properties like being yellowish in colour, being a chemical element with the symbol ‘au’, and so on.

According to the substance dualist, we can identify and contrast physical substance and mental substance using exactly the same kind of method. The central features of physical (or ‘material’) are things like having spatial extension in the world, being decomposable, and being divisible (we’ll come back to this property in a moment when we come to the argument from indivisibility). Others listed by Descartes include things like spatial location, being an object that can be sensed, passivity, being law-governed, measurability, and publicity.

Mental substance, by contrast, is a substance that does not have any of these properties. It’s not spatially extended, divisible, decomposable, or have any of the other physical properties mentioned in the last paragraph. This is because it’s not made up of the physical stuff we can see in the world around us. Rather, it’s a kind of ghostly or ethereal substance of the kind that things like souls, or minds are made of. Its properties, for Descartes, are things like being the seat of consciousness and thought, being private, being active, being indivisible, lacking spatial location or extension, and being indestructable. (Of course, in terms of this taxonomy, the example given above was a bit misleading, since both water and gold will count as physical substances).

Human persons

Us humans are made up of a combination of both of these substances. Let’s take our bodies first. Our bodies, for the substance dualist, are built out of purely physical substance. We can easily see that this is the case by noticing that it has all the kinds of properties that Descartes uses as a marker of the physical (it’s spatially extended, decomposable etc.) As a physical object, it has no problem in navigating its way around the physical world. It can pick things up, bump into things, adjust the position of things, and so on. When it engages in this kind of interaction in the physical world, according to Descartes, we should think of it as moving around just like a machine.
would. It’s propelled into its various behaviours and exerts its causal influence on things in the world according to purely physical-mechanistic laws.

This, however, is not all there is to being a human. Indeed, this kind of purely physical description of our engagement with the world seems to miss out the most important bit, which is that we also have a kind of rich awareness of what’s going on; we are conscious of the world, and of ourselves in the world, and can think about what’s happening in all sorts of ways. That is, we also have minds. These minds, we’ve already seen, are the things that account for our consciousness and thoughts. This is because according to Descartes, having consciousness and thoughts are the kinds of properties that, for the substance dualist, are distinctive of mental substance.

**Mind-body problem**

Like the other theories of mind to be discussed in these notes, substance dualism can be seen as a response to the mind-body problem. Just as a quick recap, the mind-body problem was the challenge to explain how our minds relate to our bodies. The force of this question, we saw, came from two sets of intuitions that seemed to stand in tension with each other. On the one hand, the mind and the body seem to be two very different kinds of thing, but on the other hand, they seem to be very intimately related.

Substance dualism responds to the mind-body problem by taking the first set of intuitions very seriously. The reason why it *seems* to us as if the mind and the body are very different sorts of entity is because they are! The first set of intuitions, then, have been accounted for on this view. There is, however, still a task left for the substance dualist to do. Anyone who holds this view still has to account for the second set of intuitions about how closely the mind and the body are intermingled.

As we will see in two of the arguments against substance dualism, this is really the view’s weak spot. It turns out to be very difficult to say how the mind and body interact once you say that they are made up of different substances.

**Descartes and mind-body intermingling**

Descartes himself was very aware of this task facing his theory, and unlike some of the other substance dualists we will see (occasionalists, parallelists and epiphenomenalists), give a direct answer to it. Unfortunately, his answer was not altogether convincing. His solution was to say that our minds can causally interact with our physical bodies in the same way in which we might pull a lever on a machine to set it in motion. This is not as metaphorical as you might think – for Descartes, this ‘lever’ really did exist, and was located in the pineal gland, a small endocrine gland located in the brain just above the cerebellum.

The resultant fusing of mind and body when the lever is pulled, for Descartes, is so intimate that it gives rise to certain properties that can be had by neither the mind nor the body alone. Sensations, for instance, such as pains, itches, tickles and tingles, can be ascribed only where the mind and the body are so joined; they are, strictly speaking, neither purely mental nor purely physical properties, but properties of a thing made up of the fusion of mind and body.
Any substance dualist account like this, on which the mind is taken to be causally interactive with the body, is called interactionism.

**Historical interest**

It should be kept in mind that substance dualism is very much out of favour in contemporary philosophy of mind (with a very few exceptions - see further reading). In large part, its decline in popularity came from a science-led 19th century movement to bring our understanding of the mind under a world-view that appeals only to entities we would recognise as part of the natural, or physical, world. As a part of this effort to ‘naturalise’ the mind, philosophers started looking for a way to answer the mind-body problem that avoided the substance dualist’s mysterious ‘ghosts in the machine’ (as the philosopher Gilbert Ryle called it). Some of the results of this search will be outlined in the later sections on materialist or physicalist accounts of the mind.

Because of this, substance dualism is now normally taught as part of the historical overview of the mind-body problem, and as a useful alternative with which to contrast the substance monist physicalist theories that dominate contemporary philosophy of mind.

**Property dualism**

A final quick cautionary note before we move on to the arguments for and against this view: substance dualism, which says that there are two substances in the world, should not be confused with property, or aspect, dualism, which says that there is only one kind of substance in the world, but that the substance in question has both mental and physical properties.

**Teaching tips**

- It’s likely that most students will intuitively be quite sympathetic to dualism, so this is a good place to start in a course on the philosophy of mind.
- This also means, however, that the arguments against it might be a bit of a hard-sell. It’s important, however, that the students are persuaded of the inadequacy of substance dualism if they are going to stay on board with the search for the right physicalist theory of mind for the rest of the course, so it’s worth spending a bit of time on why it can’t be right.
- As mentioned in the last section, one of the reasons substance dualism is intuitively attractive might be because it is a picture of the mind that’s pretty pervasive across various difference cultures and times, and fits in with any view on which we have non-bodily souls or spirits. As suggested in the last section, starting with these kinds of examples can work as a good way into the subject.

**Discussion questions**

- Descartes thought that only human animals had minds as well as physical-mechanistic bodies, and that other animals only had the physical body. How
plausible do you think this is? Do you think animals have mental states like beliefs, intentions, regrets or hopes?

- What do you think of the idea that our minds and our bodies are completely distinct in the way suggested by the substance dualist? Is your body just a kind of physical vehicle for your mind? Is your mind something separate that controls that vehicle like a lever on a machine?
- What do you think of the idea that our minds are made of something non-physical? Are you much persuaded by the idea that it’s made of a kind of non-physical substance? What would the substance be like?
- If you are attracted by the idea that our minds are really separate from our bodies, can you articulate the reasons why you like the idea? Is there something special about the mind that this theory would be able to explain?

Further resources

- Essential reading for this topic is Descartes’ sixth meditation from Meditations on First Philosophy (lots of different editions available, a good one is from the Cambridge Texts in the History of Philosophy series edited by John Cottingham.) Although this is a very rich text, it is also eminently approachable and by no means an unenjoyable read.
- There are very few modern day substance dualists. One exception to this is John Foster, whose defence of Cartesian dualism can be found in The Immaterial Self: A defence of the Cartesian dualist conception of the mind, New York: Routledge (1996).
1.1 Argument for substance dualism: the indivisibility argument

One of Descartes’ central arguments for substance dualism is known as the *indivisibility argument*. This argument is somewhat simpler than the other argument in favour of substance dualism to be discussed in these notes, and so might be a good place to start.

**Leibniz’s law**

The main philosophical tool that will be needed for the argument is the *indiscernibility of identicals* principle or, more familiarly, *Leibniz’s law* (named after the 18th century German philosopher and mathematician Wilhelm Gottfried Leibniz, to whom we owe this principle). Leibniz’s law is the principle that two things can be identical only if they share all the same properties. This, really, is just common sense. If two things are identical, then surely they must share all the same properties; they are one and the same thing! We might, of course, have different names for that thing, which probably explains why the question of identity would have come up in the first place. We might, for instance, ask whether Clark Kent and Superman are identical. Given that “they” really are identical, however, that man must surely have the same properties whichever name we use for him.

We should be careful to distinguish this *indiscernibility of identicals* principle, that says that things that are identical must share all of their properties, with the nearby and easily conflated principle of the *identity of indiscernibles*, which says that any two entities that have all of the same properties must be the same thing. While the former principle is uncontroversially accepted by most philosophers, the latter is somewhat less so. Fortunately for us, it is the former principle that is needed for the present argument.

**The indivisibility argument**

With Leibniz’s law in hand, we are ready to give the basic shape of the indivisibility argument as follows:

1. The body has property $x$
2. The mind does not have property $x$
3. Therefore, by Leibniz’s law, the mind and the body are not identical

I have given the argument’s schematic form here to show that the same template could be used to generate various different arguments for substance dualism – all you need to do is find a property that the body has and that the mind lacks, or vice versa. In the case of the indivisibility argument, property ‘$x$’ stands for the property of *being divisible*.

Fully spelled out, then, the indivisibility argument is:

1. The body has the property of being divisible
2. The mind does not have the property of being divisible
Therefore, by Leibniz’s law, the mind and the body are not identical.

**Divisibility of the body**

Let’s go through each of these steps in turn. What does it mean to say that the body is divisible? Gruesomely enough, this means exactly what it says: I could, if I so chose, cut my body into halves or thirds or quarters, or I could separate my left leg from the rest of my body, or, more generally, I could carve myself up any way I like.

Less dramatically, the idea here is that the body is a material object, and it is a feature of material objects that we can divide them into smaller pieces without compromising the material nature of those body-bits. Indeed, for Descartes, this kind of divisibility is one of the defining features of things made of material substance (see section 1 on substance dualism). Cut my material body into two, and what you’ll get is just two (somewhat smaller) material objects in the world. This, really, is all there is to the first premise in the argument.

**Indivisibility of the mind**

The second premise, that says that the mind is not divisible, is no trickier than the first. It just says that the mind is not the kind of thing that we could chop up into smaller pieces. After all, what would it mean to say that the mind is divisible? Can I cut my thoughts in two? Or my headache into thirds? It just doesn’t make sense to talk like this when thinking about the mental domain, in the way that it is perfectly intelligible when thinking about the body.

Descartes describes this indivisibility of the mind in terms of the mind being a simple substance, by which he merely means that it is not a substance that can be divided up into smaller proper parts. It is an indivisible whole in itself, which either exists in its entirety, or doesn’t exist at all.

**Non-identity of mind and body**

Descartes, then, takes the mind and the body to differ in the kinds of properties they have – at the very least, for all that has been said here, they differ in terms of their divisibility properties.

It is at this point in the argument that we need to make use of Leibniz’s law, the principle introduced above that if any two things are identical, then they must share all of the same properties. After all, all that we have shown so far is that the body has a property that the mind lacks – namely, that of divisibility. In order to move from here to the conclusion that the body is distinct from the mind, we must think that a difference in properties is enough to establish a difference in entity. Luckily, this is precisely the move made permissible by Leibniz’s law: given that the body and the mind do not share all of the same properties, according to that law they can’t be the same thing.

This is all there is to the indivisibility argument.
Teaching tips

- As remarked above, this is one of the simplest arguments in support of substance dualism, so a good one to ease the students into the topic.
- Because of its simplicity, it’s also a good argument to get students to develop their more general philosophical skill in handling arguments and for showing them what a syllogistic argument looks like.
- You could also use it to introduce them to the notions of validity (if all the premises are true then the conclusion must be true) and soundness (valid + the premises are true).
- You could also get them to think about how an argument of this kind of syllogistic form could be undermined – i.e. either by denying its validity, or by denying its soundness by rejecting one of the premises.
- The only slightly tricky part of this argument is Leibniz’s law. One way to help them get a grip on it could be to try to get them to come up with counterexamples to it, which will hopefully get them to think carefully about what the principle amounts to.

Discussion questions

- Once we have Leibniz’s law in hand, it looks like there might be other arguments we could generate using the same argumentative template. What kinds of properties might we appeal to? Consciousness? Object-directness/intentionality? You could have a look back at the properties that Descartes takes to distinguish mental and physical substance for inspiration.
- It seems like premise 1 is pretty robust, but what about premise 2? Is it right to say that we can’t make sense of dividing up the mind? What about people with multiple personality disorders? Or what about sleep episodes – don’t they ‘cut up’ our episodes of awareness in some sense? Are these kinds of divisibility still different enough from the divisibility of the material body for the argument from divisibility to still go through?

Further resources

- The idea that sleep disrupts our mental continuity was given by John Locke in *An Essay Concerning Human Understanding* bk. II, Ch.1, §10. (Online pdf available here: [http://www2.hn.psu.edu/faculty/jmanis/locke/humanund.pdf](http://www2.hn.psu.edu/faculty/jmanis/locke/humanund.pdf))
- The view that our mental lives are a simple unity is challenged by David Hume’s ‘bundle theory’ of the self, according to which the ‘selves’ that we introspect are a bundle of ideas and impressions rather than a unified entity. His most well known formulation of this theory is in *A Treatise of Human Nature*, bk. 1, pt.4, §6. (The relevant section is available online here: [http://www.anselm.edu/homepage/dbanach/pi.htm](http://www.anselm.edu/homepage/dbanach/pi.htm))
• There is a wealth of material available on disassociative and multiple personality disorders, but here are a few popular articles and medical resources that might be helpful as a starting point:
  o http://psychology.about.com/od/psychotherapy/f/psychvspsych.htm
  o http://www.guardian.co.uk/lifeandstyle/2011/sep/30/kim-noble-woman-with-100-personalities
  o http://www.psychologytoday.com/conditions/dissociative-identity-disorder-multiple-personality-disorder
  o http://www.mind.org.uk/mental_health_a-z/8039_dissociative_disorders
  o http://www.nami.org/Content/ContentGroups/Helpline1/Dissociative_Identity_Disorder_(formerly_Multiple_Personality_Disorder).htm

• Stanford Encyclopedia’s online resource is often a good starting place with a new topic – it provides detailed topic overviews that are serious-minded and somewhat advanced, but that are nevertheless fairly accessible and clearly written. They are also a great resource for further reading suggestions, as each topic is accompanied by an extensive topic bibliography. The entry on Leibniz’s law is here: http://plato.stanford.edu/entries/identity-indiscernible/
1.2 Argument for substance dualism: the conceivability argument

Imaginability of mind-body separation

The most well-known argument for dualism – the so-called conceivability argument – begins with the thought that if I set myself to the task it is fairly easy to imagine waking up one morning and finding to my great surprise that my body has disappeared. I can, in my imagined scene, seem to see the room around me and the crumpled bedclothes beneath me. Everything seems to be in place, but when I look down, the familiar assembly of arms, knees and toes I’m used to seeing when I look downwards just isn’t there. This might, of course, all seem a bit fanciful. What is important for the argument from conceivability, however, is not that this scene is in any way plausible or likely to happen, but only that it is imaginable.

This might, at first, not seem terribly impressive. Can’t we imagine all sorts of things? Why should we be interested in what things we can imagine? Well, it turns out that there are some things that we don’t seem to be able to imagine. Try putting your mind, for instance, to the attempt to imagine a second scene in which you wake up one morning to find that \(2+2\) no longer equals 4, or again a third in which you awake to find that the inside corners of all of the triangles in the world no longer add up to 180 degrees. No matter how hard you try, imagining these things just seems to be impossible. Unlike the first scenario, these last two cases don’t seem to provide a coherent enough narrative to be imaginable as a scene; what would it be to imagine \(2+2\) equaling something other than 4, or to imagine a triangle that didn’t have its internal vertices add up to 180 degrees? We can, of course, imagine ourselves making mistakes about these matters, but that is a very different thing to imagining a world in which they are actually true.

The first step in the argument, then, is to establish that unlike these other cases, we can imagine a mind existing without a body, and so – importantly – imagine a mind existing without a brain. How do we get from here to an argument for substance dualism?

Conceivability-entails-possibility principle

Well, imagining this scene is a way of conceiving of it (indeed, such imaginative visualisation, we might think, is really the main way that we usually conceive of things). And it is the conceivability of this scenario that tells us something important. It tells us that there is no internal contradiction to the imagined state of affairs, that there is no inconsistency in thinking that this is a way that the world might have been, even if it did not actually turn out that way.

More precisely put, the next move in the argument employs the principle that the conceivability of a given state of affairs entails its metaphysical possibility, where ‘metaphysical’ possibility can just be thought of as possible ways in which the world might have been if the natural laws of the world had turned out differently. It might be helpful to contrast this ‘metaphysical’ possibility with the more limited ‘physical’ possibility, which marks the ways things might have turned out given the natural laws that really do govern our world. It is, for instance, metaphysically but not physically
possible that a human being could jump over the channel, or that gravity could only
be operative on alternate days. (It should be noted that this conceivability-entails-
possibility principle is far from uncontroversially accepted in the literature – see
further resources.)

Here, then, is state of the argument so far: we have seen that a bodiless (and so
brainless) mind is conceivable, and have found that this conceivability reveals such a
state of affairs to be metaphysically possible. So far, however, this only shows that
this is a way that the world might have been. But we are trying to argue for substance
dualism, which is a position that says that as things really are the mind is not identical
to the brain. How does the metaphysical possibility of a mind existing without a brain
show us that the mind is actually a different thing to the brain?

Necessity of true identities

The place to look in response to this question is to what kinds of metaphysical
possibilities are going to hold for identity-statements, (or statements that say that one
thing is identical to another thing). To get ahead of myself, the answer is going to be
that when something really is identical to another thing, then their separation can’t be
metaphysically possible. The fact that a mind-brain separation is metaphysically
possible, then, shows that those two things can’t really be identical. And if this is
right, then substance dualism is correct. But let’s slow down a bit and see how we get
to that point. Why should we think that when things are really identical, it’s not
metaphysically possible that they could come apart?

Perhaps an example will help to illustrate why this is. ‘Dhaulagiri’ is the Nepali name
for the mountain that we know more familiarly as ‘Mt Everest’. Here, then, we have
an identity-statement that we probably have a better handle on than the mind-brain
identity, and which might help to get us going on this tricky part of the argument.
The identity-statement in question is the claim that Dhaulagiri is identical to Mt
Everest. The two are just one and the same thing. When you climb one of them,
then, you also thereby climb the other, when you photograph one you also photograph
the other. You couldn’t do otherwise given that they are the same thing.

We might, at this point, already have inkling as to why it couldn’t be metaphysically
possible that a true identity could come apart. For consider now the suggestion that
Dhaulagiri could have been distinct from Mt Everest. What does this suggestion
amount to? Well, given that they are actually just the same thing, the suggestion
amounts to the idea that a certain mountain on the Nepal-Tibet border might have
been non-identical to itself, an obviously nonsensical suggestion. For how could
something fail to be the same thing as itself? If two things really are one and the
same, then, they must be one and the same in all metaphysically possible scenarios.
Changing the natural laws of the world (which, remember, is what we’re testing when
we ask about metaphysical possibilities) wouldn’t change the logical truism that
everything is identical to itself.

The same thing, then, applies to the supposed identity of the mind and the brain. To
say, along with the substance dualist’s opponent, that they are identical is to say that
they are one and the same thing. But if they really are one and the same thing, then it
makes no sense to say that they could be separated in other metaphysically possible
scenarios – how could something be separated from itself, even if the natural laws of the world were different? The suggestion just makes no sense. Their identity, if it really is an identity, must therefore hold in all metaphysically possible worlds. But we have already said that the conceivability of a disembodied mind shows that it is metaphysically possible that a mind could exist independently of a body. The mind and the brain, therefore, can’t really be identical. Substance dualism must be correct.

The conceivability argument

Let’s summarise the argument just given in a more concise form:

(1) It is conceivable that a mind could exist without a body
(2) Therefore it is possible that a mind could exist without a body
(3) Therefore the mind is distinct from the body (substance dualism)

The transition from (1) to (2) is given by the conceivability-entails-possibility principle, and the transition from (2) to (3) is given by the idea that all true identities are metaphysically necessary, or the thought that it cannot be metaphysically possible that they could come apart.

Zombies

This conceivability argument, which had its earliest formulation in Descartes, is a powerful one; indeed it foreshadowed an argument that is still used today against physicalism (the dominant opposing view to dualism) in its various forms. This is the so-called zombie argument which, like Descartes’ argument, works by persuading us of the conceivability of a separation of the mind and body. The only difference is that instead of conceiving of a minded subject who doesn’t have a body, the zombie argument works the other way around: it works by inviting us to conceive of a mindless bodily subject – namely, a zombie!

A zombie should be thought of as a perfect physical replica of a human being. This is so both in terms of its outwards behaviours – it goes about the world reacting to things in just the same way that we do – and in terms of its physical composition. Most importantly, crack open its head, and you will find a brain that is in every respect exactly like a human brain.

With these zombies in hand, we can construct the argument in exactly the same form as the original conceivability argument:

(4) Zombies are conceivable (a body without a mind is conceivable)
(5) Therefore zombies are possible (a body without a mind is possible)
(6) Therefore the mind is distinct from the body (substance dualism)

The transitions between these moves in the argument parallel the transitions in the original: the move from (4) to (5) appeals to the conceivability-entails-possibility principle, and the move from (5) to (6) works by recognising that if it’s possible for a seeming identity to come apart, then it can’t really be a true identity. If the mind really is just one and the same thing as the brain, that is to say, then zombies shouldn’t be possible. This is because saying that zombies are metaphysically possible at the
same time as saying that the mind is identical to the brain is just the same thing as saying that it is metaphysically possible for one thing (the mind/brain) to both be present and not to be present at the same time, but this is clearly impossible, even in merely metaphysically possible worlds!

I’ve given the zombie argument here using the term ‘body’ throughout, but this is really just a placeholder for whichever sort of physical state or property a particular version of physicalism identifies with the mind. As we will see in later sections, for example, one version of physicalism thinks that mental states are really nothing more than behavioural dispositions. A zombie argument against this so-called behaviourist version of substance physicalism, then, can be constructed by replacing mention of ‘a body’ with ‘behavioural dispositions’, and likewise for other versions of reductive physicalism.

**Teaching tips**

- There are several tricky steps to this argument. The necessity of true identities, for instance, is a difficult idea to get a grip on so worth spending some time on. Using lots of examples will help.
- The conceivable-implies-possibility principle is also a little slippery. In particular, students might struggle to grasp that there is anything that is not conceivable. Having a few inconceivable examples up your sleeve might help here (e.g. a square circle, parallel lines that meet, something that is both red all over and green all over, an object that both exists and doesn’t exist at the same time, etc. – anything that violates a logical truth will do.)
- You could come back to the zombie argument when you get to reductive physicalism, and try to get students to construct their own zombie arguments against specific forms of reductive physicalism.

**Discussion questions**

- Can you think of some other examples of scenarios that are metaphysically possible but not physically possible?
- Can you think of your own example of a true identity? Can you explain to a partner why it is metaphysically necessary (i.e. can’t be a non-identity in any metaphysically possible world?)?
- Do you find the conceivable-implies argument and/or the zombie argument persuasive? If so, are you convinced of the truth of substance dualism? If not, can you explain which bit of the argument(s) you think don’t work?

**Further resources**

- For some first-hand versions of the zombie argument, see Kripke’s 1972 lectures, *Naming and Necessity*, Cambridge MA: Harvard University Press, (1972/1980) (though he does not use the ‘zombie’ terminology), David

- The Stanford Encyclopedia entry on zombies and zombie arguments can be found here: [http://plato.stanford.edu/entries/zombies/#3](http://plato.stanford.edu/entries/zombies/#3)
- For a rather detailed exploration of the conceivability-entails-possibility principle, see David Chalmers’ ‘Does Conceivability Entail Possibility?’ in T. Gendler & J. Hawthorne, eds) *Conceivability and Possibility* (Oxford University Press, 2002), pp.145-200. The second half is a little dense, but the opening sections are more accessible and provide a helpful and in-depth discussion of the principle. An online version of the paper is available on Chalmers’ personal website: [http://consc.net/papers/conceivability.html](http://consc.net/papers/conceivability.html)
- The Gendler and Hawthorne volume just mentioned is a collection papers on the topic of whether conceivability entails possibility – but perhaps worth mentioning that some of these are a little technical. The book has a clear and detailed introduction, which can be found here: [http://fds.oup.com/www.oup.co.uk/pdf/0-19-825090-8.pdf](http://fds.oup.com/www.oup.co.uk/pdf/0-19-825090-8.pdf)
1.3 Argument against substance dualism: the interaction problem

The first problem that I will set out for the substance dualist is the interaction problem, a problem often taken to be fatal for substance dualism. It was, as far as we know, first put forward by Princess Elizabeth of Bohemia in a letter to Descartes. (See further resources).

The problem emerges from the idea that the mental and the physical realm are causally interactive. This is just to say that things that happen in the mind sometimes cause physical things to happen, and in return that things that happen in the physical body or physical world sometimes cause mental events to take place.

That this kind of interaction of the mind and the body is just a common sense idea can be brought out with a pair of examples illustrating this causal interplay in both directions. In the first direction, the suggestion is that the mind can cause physical events to happen in the body. This just seems obvious: to give a recent example, a moment ago my desire to appease my sweet-tooth (a mental state) caused me to walk to the kitchen and get some chocolate (a physical event involving complex bodily movements). Likewise, it seems undeniable that causation can also go in the opposite direction. A brutal toe-stubbing (physical) can naturally be said to cause feelings of pain and self-pity (mental).

The interaction problem

The interaction problem amounts to the challenge to the substance dualist to explain how such causal interaction could take place between substances as radically different – on the substance dualist’s account – as the mind and the body. How is it possible? It is deeply mysterious, if the mind really is made up of a non-spatially-extended substance, how it could exert causal influence over a spatially located object in the world like one’s body. How can a non-located entity causally affect a located one? Or, equally perplexing, how could an ordinary physical object like a body serve to cause changes in a non-spatial ghostly substance like the mind? Where would these causal exchanges take place? It certainly doesn’t look anything like our normal picture of causation on which, for example, I kick a ball and it moves off with predictable speed and direction. And yet, the thought that there is such causal interaction between the mind and the body is very difficult to resist.

The substance dualist is left with an explanatory gap which has yet to be filled in satisfactorily; even Descartes had little to say in response to it. Notice that it is really this failure on the part of the substance dualist rather than the argument itself that gives the interaction problem its force. After all, we rarely take the fact that we don’t yet know how to explain something as decisive reason to reject the phenomena that gave rise to the explanatory question in the first place – scientific disciplines would hardly get very far if we did! It is, then, the fact, that the explanatory gap in this case is so apparently difficult to fill in that really gives the interaction problem its teeth.

A substance dualist response

Of course, even if the substance dualist cannot find a way to explain how it is that the mind and the body could causally interact, there is a second kind of response
available. The substance dualist might simply deny the premise that the mind and the body really do interact in the ways that we suppose. There have been a number of such substance dualist responses (none of which, it has to be said, have proved to be very convincing.) I will outline the two best known such responses here, both of which attempt to solve the interaction problem by denying the intuition of mind-body interaction.

**Occasionalism**

The first of these is called *occasionalism*, a view propounded by one of Descartes’s followers, Nicolas Malebranche. Broadly speaking, occasionalism is the view that God is the only causal power, and so that everything that happens in the world is caused by Him directly. More specifically, the occasionalist responds to the interaction problem by holding it to be a mistake to suppose that the mind and the body can and actually do causally influence each other. In fact, on every occasion on which it seems as if there is such an interaction, God is intervening so as to cause a mental experience appropriate to the physical event, or vice versa. By denying the veridicality of the appearance of mind-body causation, the occasionalist disintegrates the interaction problem’s challenge to substance dualism; we don’t have to explain how mind and body could causally interact because there is nothing here to be explained – in actual fact, despite appearances they don’t really interact at all.

**Parallelism**

A second response that uses the same kind of strategy to answer the interaction problem is called *parallelism*. Like the occasionalist, the parallelist holds that the mental realm and the physical realm are, in fact, causally insulated from one another. Unlike the occasionalist, however, the parallelist does not think that God intervenes on each occasion to ensure that the two realms appear to be in line with each other. According to the parallelist, rather, both realms proceed in their course in absolute isolation from each other, but nevertheless always end up matching up so that to some it may seem as if they interact.

There is, of course, something a bit suspicious about this supposed coincidence, which the strongest form of this view attempts to explain by appeal to a higher power. Leibniz’s *pre-established harmony* version of parallelism, for instance, holds that the mental and the physical realms appear to be causally linked only because God has already arranged that the two will progress in perfect coordination. A helpful way to think of this, perhaps, is that it is a bit like God pressing ‘play’ on a silent film and an independently recorded soundtrack at precisely the same time.

**Teaching tips**

- As noted above, the force of the interaction problem really comes from the substance dualist’s failure to come up with a satisfying response to the explanatory challenge. One way to get students to see the power of this problem is to get them brainstorming on the part of the substance dualist, each time trying to show that their suggested answer won’t work. This way they will get to feel the frustration of the substance dualist for themselves, and hopefully be moved by the force of the challenge.
Discussion questions

- How convincing do you find the interaction problem? Aren’t there other distinct kinds of things that can causally interact despite their difference in kind? (E.g. what about gasses interacting with solids or liquids? Response – these are nevertheless all part of a spatially extended world, and it is the non-spatial nature of the mind that generates the real mystery in the case of the interaction problem.)

- This problem could be a platform to start a discussion about what we mean by causation. It seems like, if this is going to be a problem, we have to think of causation as the kind of thing that happens when objects push other objects around the world. But are there other ways of understanding what causation is that doesn’t depend on anything to do with the nature of the objects involved? E.g. what about a counterfactual story of causation, where A causes B if and only if B wouldn’t have happened if A hadn’t happened? Or a probabilistic story of causation, on which As cause Bs if and only if As raise the probability of Bs occurring? Do we even really have an intuitive grip on what causation is beyond constant conjunction (i.e. one thing always seeming to happen after another thing)?

- Even if, like the occasionalist or the parallelist, you deny the causal interactivity between the mind and the body, it still looks like you would have to explain why we have such a profoundly held conviction that the mind and the body do causally impact on each other. (Such an explanation is called an ‘error theory’, and is required whenever an argument suggests that our everyday intuitions are wildly off the mark). What could the occasionalist’s or the parallelist’s error theory be? Do you find them convincing?

Further resources

- The correspondence between Descartes and Princess Elizabeth of Bohemia is great fun to read – http://www.earlymoderntexts.com/pdf/descelis.pdf. The interaction problem comes up in the first letter, and the challenge as she puts it is still a very clear and helpful way to think of it.

- The Stanford Encyclopedia entry on mental causation is useful for a broader overview of related issues: http://plato.stanford.edu/entries/mental-causation/

- For more on Leibniz’s parallelism see the Stanford entry on his philosophy of mind: http://plato.stanford.edu/entries/leibniz-mind/. The Leibniz itself can be a little opaque, especially since some of his writings on the topic are bound up with other fairly obscure parts of his metaphysics. One of the more perspicuous presentations of it is in his correspondence to the philosopher Samuel Clark, in particular his fifth letter in this collection: http://www.earlymoderntexts.com/pdf/leibclar.pdf.

- For Malebranche’s occasionalist theory, see §§8-15 of dialogue 4 here: http://www.earlymoderntexts.com/pdf/maledial.pdf

- See also the chapter on this topic in Tim Crane’s introductory book to the philosophy of mind, Elements of Mind, Oxford: OUP, 2001.
1.4 Argument against substance dualism: the argument from the causal closure of the physical

A second argument often made against substance dualism – *the argument from the causal closure of the physical* – is a close relation of the argument from the last section insofar as they both pivot on the intuition of causal contact between the mental and physical domains. Unlike the interaction problem, however, which was motivated by the intuition of mind-body causation going in both directions, the argument from the causal closure of the physical is only interested in a single direction of causality: namely, the apparent causal influence that our mental lives have over our bodies and the things that we do with them. We have the intuition that the mind can make an actual difference to what happens in the physical world.

**Mind to body causation**

The intuitive compulsion to think that our mental states and events have this kind of causal influence over our bodies is, we saw in the last section, so strong as to be almost irresistible. If asked to explain why I just reached my arm towards my coffee cup, my answer is likely to mention something to do with my state of mind just before the action occurred – my desire for a sip of coffee, for instance, or a fatigued feeling paired with my belief that coffee-drinking is just the thing to shake me out of it. In fact, put this way it just looks like an obvious everyday truth that things that happen in my mental life are the kinds of things that often genuinely influence physiological changes that take place in my body (a movement of the relevant limb, etc.) The argument from the causal closure of the physical uses this intuitively compelling causal claim together with a pair of equally plausible-sounding additional premises to generate an apparent contradiction in the substance dualist’s account.

**The argument from the causal closure of the physical**

Here is the argument from the causal closure of the physical:

(1) Mental states often, or systematically, cause physical events
(2) Every physical event has a sufficient physical cause
(3) Physical events do not systematically have more than one sufficient cause

Therefore mental causes must be identical to physical causes

**The premises**

Let’s consider each premise in turn.

(1) is just a statement of the intuition about mental causality discussed both in relation to the interaction problem, and earlier in this section. Enough, I hope, has been said to motivate this premise. The only thing that I have added here is that our mental states cause physical events *often*, or *systematically*. All I mean by this is that there is a kind of regularity between the kinds of mental states I have and the kinds of physical effects they cause (coffee-desire in the right kinds of contexts, for instance, systematically cause arm-reachings towards coffee cups). When we get to premise
(3) it will, I hope, become clear why I have had to add this in order to get a tension
going between the premises.

Premise (2) is a statement of the thesis of the causal closure of the physical that gives
this argument its name. There are various ways in which we might have chosen to
formulate this thesis. A strong formulation, for instance, would have it that every
physical event has a unique physical cause. The way I have formulated it in the above
argument is a little weaker than this; it says only that for every physical event there is
at least one sufficient physical event that has caused it. This, really, is just the
thought that for any given physical event (the event of my fingertips reaching the
coffee cup, for example) there will be a fully specifiable physical explanation that al.
Reaching for a coffee cup, for instance, can be traced back to neuron activity in the
cerebellum and motor cortex. These neurons fire to the neuromuscular junction
causing contraction in certain muscles and relaxation in the opposing muscles in the
arm and hand.

We could, according to (2), describe all of the physical happenings in the world
(including our bodily movements) in this way, using purely physical language. If that
is right, then for every physical event we are sure to find a physical cause – to think
otherwise, the thought is, would be something like believing in telekenesis.

There are also somewhat more theory-laden arguments than this that we could have
appealed to to motivate the second premise. The arguments I have in mind rely on the
fundamental scientific principle of energy conservation which, roughly speaking, says
that the total amount of energy within a closed physical system can’t ever go up or
down, even if it can move around the system. If this is right (and it is almost
universally accepted) then any physical event must be wholly describable as an
exchange of energy occurring entirely within the physical system in question. This is
probably more than would be needed to motivate (2) at this level so I won’t pursue
this line of thought here, but see further resources below.

Premise (3) says that physical events are not systematically overdetermined, or, to put
it another way, that they do not typically have more than one sufficient cause. Why
should we accept this premise? Surely we can think of examples in which a physical
event does seem to be overdetermined in this way? To take a stock example, imagine
the scene of an execution by firing squad in which each shooter fires a bullet at their
common target at exactly the same time. Suppose that all of the bullets enter the body
at precisely the same instant, and moreover, that each shot would have been fatal by
itself. The target’s death, in this case, is overdetermined – it is the result of multiple
independently sufficient causes. It doesn’t seem problematic in this case to say that
no single shooter caused the victim’s death, even if each would have done if the
others had not been there. Why, then, can’t we think of the causes of our bodily
physical events in the same way? That is, why can’t we just say that they are
systematically caused by both the physical causes mentioned in premise 2 and the
mental causes mentioned in premise 1 in just the same way that the target’s death was
caused by all of the shooters at once?

There is, as it turns out, a wealth of reasons to reject postulating this kind of
systematic overdetermination. One of these is that if each of these causes really were
sufficient for its effect, then it would be a deeply puzzling coincidence that the two
causes nonetheless always happen in tandem. If my desire for coffee is really
sufficient to cause my reaching for the cup, then why doesn’t it sometimes do so by
itself? Why, come to that, aren’t the purely physical causes sometimes efficacious in
the absence of any mental cause?

A second reason to be suspicious of systematic overdetermination is that it generates a
picture on which there are always more causes than are needed. Isn’t it a bit strange to
say that mental states and events are causally efficacious, even though when it comes
down to it there really is no causal work left for them to do? Aren’t they just
redundant? A third, and related, consideration takes off from the principle known as
Ockham’s razor, which says that we should not postulate more entities than needed
when building our theories, or that the simplest of two competing theories should
always be preferred. To postulate lots of additional causes in a widespread and
systematic way would be to seriously violate this principle.

With all these reasons for accepting it, then, (3)’s rejection of overdetermination looks
itself to be suspiciously overdetermined!

Identity of mental and physical causes

So long as we accept premises (1) to (3) it looks like the conclusion must follow that
the mental causes of physical events must be identical to those events’ physical
causes. This seems to be the only way to reconcile all three premises, each of which –
as we have seen – seem to be in good standing on their own; if we are in agreement
with premises (2) and (3) that all physical events have at least and at most one cause
which is itself physical, then the only way to hold on to the powerful intuition that
mental events regularly cause physical events is to say that those mental events are
just the same thing as those physical events. The neuronal activity to which I can
trace back the physical causal history of my arm-reaching, for instance, must just be
one and the same thing as the mental event of desiring coffee that featured in my
commonsense explanation of that action. To say this much, however, is to deny
substance dualism.

A substance dualist response: epiphenomenalism

A response that has sometimes been given to this argument on behalf of the substance
dualist comes from a view called epiphenomenalism, which uses the strategy of
simply denying that mental states and events ever cause any physical effects.
According to the epiphenomenalist, while it is correct to say that physical things can
cause mental happenings, those mental events cannot in turn assert causal influence
over anything physical; they are causally inefficacious ‘epiphenomena’ that are
caused but are not themselves causes of anything. Epiphenomenalists thus neutralize
the argument from the causal closure of the physical by denying premise (1).

This view is, in a way, more palatable than occasionalism and parallelism (from the
last section) in that it denies only one half of the mind-body causality intuition where
occasionalists and parallelists deny it in both directions. Having said that, however,
this restraint on the part of the epiphenomenalist also makes it less satisfying than the
other two responses in another respect. By denying that mind-body causation is
possible in either direction, at least occasionalists and parallelists have in hand a
response to both the argument from the causal closure of the physical and the interaction problem. Epiphenomenalists, by contrast, are still no better off than their fellow interactionist substance dualists in the face of the causal interaction problem. In fact, epiphenomenalists seem to face a revived version of that problem; if it is right that mental occurrences can have physical causes then it is even more mysterious why they can’t in turn causally influence the physical. Moreover, unlike occasionalism and parallelism, epiphenomenalism itself does not provide us with any explanation of why it seems so strongly to us as if there are sometimes mental causes of physical effects.

The causal closure of the physical and the interaction problem

The commonalities and differences between the arguments make the interaction problem and the argument from the causal closure of the physical naturally complementing arguments to teach alongside each other; while they both exploit the same family of intuitions about the causal interactions between the mental and the physical realms, they are arguments of very different kinds. The interaction problem is a kind of ‘brute’ argument that works by simply confronting the substance dualist with a seemingly unanswerable explanatory challenge. While the force of the argument is immediately felt, then, it can nonetheless feel like a less-than-satisfying victory over the substance dualist. By contrast, although the argument from the causal closure of the physical relies on a number of premises that require a bit of working out, the syllogistic form of the argument can make its conclusion feel more gratifyingly ‘earned’ than that of the interaction problem. Its conclusion is also a bit sturdier – whoever accepts the premises of this argument is seemingly obliged to conclude that substance dualism is mistaken.

Teaching tips

- This is another nice argument for demonstrating the notions of soundness and validity, and showing what you would need to do if you wanted to reject the conclusion of the argument – i.e. either argue that it’s invalid, or argue that it’s unsound by denying one of the premises.
- This argument has a few steps that rely on things other than pure intuition, so will be a harder sell to students than the more intuition based arguments (like the interaction problem). For the moves in the argument that are a bit more theoretical (e.g. the problems with systematic overdetermination) use lots of examples to show the students the attraction of these kinds of theoretical virtues.

Discussion questions

- Some responses to the causal closure argument have focused on whether or not we have a sharp enough notion of ‘the physical’ to know what a thesis like that of the causal closure of the physical would really amount to. Should it only include entities allowed into our current best physical theories? But how can we be sure that our best theories capture everything there really is to capture in a physical theory? Should we instead be talking about ‘the physical’ as a domain that includes all of the entities in some future or ideal physical theory? But then, do we really know what we’re talking about?
Another way of pushing back against this argument is to refuse to accept premise (3)’s rejection of systematic overdetermination. Although I gave a number of arguments above, none were decisive – they were based, rather on considerations of theoretical virtues (e.g. Ockham’s razor), or on inferences to the best explanation. How attractive a response would it be to say that physical events are systematically overdetermined?

How might a parallelist or an occasionalist respond to the argument from the causal closure of the physical?

If epiphenomenalism was correct, would there be a problem in even stating it as a theory? (i.e., wouldn’t the fact of our talking about mental processes count as a physical effect of mental causes?).

By saying that mental states and events are causally inert, the epiphenomenalist denies *both* that they cause physical events but also that they cause other mental events. How plausible is this claim? Can we make sense of a view on which one thought can’t cause another?

Further resources


Also, Mills’ ‘Interactionism and Overdetermination’ in *American Philosophical Quarterly*, 33 (1996), pp. 105–17 for another attempt to make overdetermination palatable, and for a clear statement of the problem of overdetermination.

1.5 Argument against substance dualism: the problem of other minds

Metaphysical vs. epistemological problems

So far the problems we've looked at for substance dualism are metaphysical; they work by revealing that if substance dualism turned out to be correct then certain problems would arise in accounting for the way we take the world to be, or for our understanding of the structure of reality. In this section we'll look at an epistemological problem for substance dualism, or problems to do with what we can know if the substance dualist is right. After all, even if the mind-body problem is itself a metaphysical question, if a particular response to it has unacceptable epistemological implications, then that's still going to count as a point against that view.

More specifically, we'll look at the epistemological problem of other minds for substance dualism.

A problem family

First, a quick cautionary note. Although it is often called the problem of other minds, there are in fact a handful of related problems which are far from unique to substance dualism. Even if mental states are not distinct from physical states, for instance, there is still a conceptual question about how we could possibly develop mental concepts that can be applied both first- and third-personally given that we seem to learn about them and how to apply them from our own case. Likewise, a non-dualist will still have to explain how it is that we come to know about others’ mental states, given that they do not seem to be the kinds of things that we can directly perceive with our senses. Moreover, even a non-dualist has to give us an account of how we should deal with the possibility of actors who are only pretending to have the mental states that they present themselves as having. Don’t these pretenders give us reason to be suspicious about the claim that we can ever know what mental state someone else is in?

I mention these problems to put them to one side – in this section I will consider a problem of other minds that is specific to the substance dualist; roughly, the problem that if the mind and the body are really distinct then given that I only ever encounter bodies in the world I have no reason to think that there are any minds in existence other than my own.

The substance dualist problem of other minds

This argument can be formulated as a reductio ad absurdum, or as a set of premises which, if true, would lead to an absurd result (or, given an additional premise asserting the denial of the absurdity in question, to a contradiction).

In this case, the reductio would run as follows. If substance dualism is correct, then the mind is composed of a non-spatially extended, non-physical substance (so, non-physically detectable substance), which could conceivably exist
independently of a body. This means that if substance dualism is correct, then there is no way of knowing about the existence of mental states other than one’s own – after all, if the mind and the body could conceivably be separated then nothing about our interactions with others’ bodies is going to count as a reliable guide to the existence or non-existence of minds or of specific mental states.

From this point in the argument, there are actually not one but two potential absurdities to be drawn. The first is that if substance dualism is correct, then we have no way of knowing that there are minds other than one’s own. The second is that if substance dualism is correct, then we don’t have any way of knowing which mental states anyone else is undergoing.

But surely these are both obviously false! Surely we do know that we are not alone in this world, or that other minded people exist; to deny this is to accept that – for all I know – the position of solipsism, which says that I alone exist in the universe, is correct. And likewise, surely we do sometimes know what kind of mental state someone else is in. When I see you writhing around on the floor clutching a gushing wound in your side, I don’t have to ask before concluding that you are in pain.

Here, then, is the reductio ad absurdum: if substance dualism is correct, then we will have to accept the absurd results that we don’t know if there are any other minds in the world, and that even if we did know this we would still never know which mental states other people are undergoing.

Teaching tips

- This is a great opportunity to introduce students to a new type of argument – the reductio ad absurdum. It might be good to have a few very simple toy examples to hand to show them exactly how the argument-type works, before moving on to the problem discussed here. You could also get them to try to construct a few reductios of their own.
- This argument is somewhat less forceful than the other arguments against substance dualism already discussed, so should probably be the one to skip if there is a choice between them to be made. The reason I include it is that it would give the students a chance to think about an epistemological argument as well as metaphysical ones, and to start exploring issues about other minds.

Discussion questions

- So long as the substance dualist accepts mind-body causation (i.e. is not an occasionalist, parallelist or epiphenomenalist), is there a way for the substance dualist to deny that their view entails that we can’t know what mental state someone is in by observing their physical behaviour? Given, for instance, that Descartes thinks that mental processes cause physical events through causal exchanges that take place in the pineal gland in a law-like and systematic way, couldn’t he say that observation of physical behaviour gives us a reliable way to infer what mental states someone is
in? Would this also work in response to the second way of putting the problem, i.e. that we have no way of knowing whether someone is minded?

- Why do we call this an epistemological problem rather than a metaphysical one? (Note that all of the premises and the conclusion are formulated in terms of what we can know, and not what is the case. For instance, one of the results is not that solipsism is correct, but that for all I know solipsism is correct, etc.)
- What is so wrong with solipsism? Or the idea that we don’t really know which mental states others are undergoing? After all, the only thing going for those premises are really our intuitions about those claims, and surely mere intuitions are the first thing to be given up in the face of philosophical reasoning?
  - Is this line of thinking encouraged by the conceivability of scenarios like Descartes’ evil demon, or the matrix? Don’t they show that I really might be the only thing that exists?
- On the other hand, isn’t the thought that we know we’re in a world with other minded people in which we sometimes know what they’re thinking and feeling pretty fundamental to our world view? Surely, then, the burden of proof lies with the substance dualist to show us that we have good reason to question it?

Further resources

- The problem discussed here is far from the most common out of the ‘family’ of other-minds problems mentioned above, so be wary of indiscriminate searching for material on the problem of other minds discussed here. Stanford encyclopedia has a good entry on this topic, that should help with disambiguation between the various related problems – http://plato.stanford.edu/entries/other-minds/
2. Reductive materialism

Broadly speaking, materialism is the view that everything that exists is fundamentally material in nature. The label ‘materialism’ is often used interchangeably with the more modern-sounding ‘physicalism’, and I will do the same here. The difference between them is that physicalism refines the central claim of materialism (that everything is fundamentally made of matter) to the claim that everything is fundamentally physical. This difference, however, will not matter for our purposes.

You can be a materialist about lots of different things. You might, for instance be a materialist about biology if you think that all biological entities in the world are ultimately material in nature, or that as long as you fixed all of the physical facts about the world, the biological facts would be fixed too.

Materialists in the philosophy of mind are – not very surprisingly! – materialists about the mind. Their claim is that all mental phenomena are fundamentally material or physical in nature. From now on I’ll use ‘materialism’ and ‘physicalism’ just to mean materialism or physicalism about the mind.

Supervenience

This definition, however, is not going to be very helpful until we get clearer on what it means to say that all mental phenomena are fundamentally material. There are a few different things it could mean, some weaker and some stronger. One of the stronger things it could mean is that all mental phenomena can be reduced to physical ones – I’ll come to that in a moment. However, even those that go in for this full-blooded, or reductive, materialism also accept a related weaker claim. This weaker claim says that mental phenomena essentially depend on, or supervene on, physical ones. Before explaining what a reduction of the mental to the physical would be, then, I’ll first explain what it means for the mental to supervene on the physical. (We’ll see later that it will be useful to already have the notion of supervenience in hand when we come to non-reductive physicalist theories).

Supervenience is a kind of relation that holds between different kinds of properties. When we say that one kind of property supervenes on another, what this means is that there can be no difference in the first kind of property without there also being a difference in the second. Equivalently, we could say that if there are two states of affairs that are exactly alike in all of the second kinds of properties, then they will also be exactly alike in the first kind of property. It’s important to note that supervenience is a relation that only goes in one direction – if one property-kind supervenes on a second, then there can be no change in the first kind of property without a change in the second, but we still haven’t ruled out that there can be a difference in the second property-kind without there being a change in the first.

An example would probably be helpful here. I’ll use the example of the property of softness. Suppose that I have an indescribably soft marshmallow in my hand. It’s clear that in some sense the softness of the marshmallow depends on its physical make-up – it is, that is to say, surely facts about the way it’s physically constituted in virtue of which it is so soft. With the philosophical tool of supervenience in hand, however, we can say something more precise than this. Namely, we can say that the
marshmallow couldn’t differ in how soft it was unless it also differed physically from
the one I have in my hand now. In other words: as long as any two marshmallows are
exactly the same in every one of their physical features, then will have to also be
exactly the same in their level of softness. As noted above, however, this dependence
only goes one way. There’s no reason why we couldn’t swap sugar for artificial
sweetener in one of the marshmallows, but still get precisely the same level of
softness as in the original sugar-marshmallow. What this shows is that softness is a
property that supervenes on physical properties, but those physical properties (e.g. the
crystalline structure of the marshmallow’s make up) don’t in turn supervene on the
softness ones.

Materialists of all stripes are committed to the claim that mental properties supervene
on physical properties. There can be no difference in what mental properties I have
(like being in pain, or remembering the beach at Hunstanton, etc.) without a
difference in my physical properties (the physical composition of my brain, for
instance). An intriguing upshot of this, and one which might provide the basis for a
class exercise, is that if I had a twin who was exactly physically alike to me in every
respect, then if physicalism is correct she would also be exactly like me in all of her
mental properties.

Reduction: explanatory and ontological

Reductive materialists, like all materialists, accept the supervenience of the mental on
the physical. But they also go in for something stronger than supervenience. They
claim that mental properties can be reduced to physical properties. What does it mean
for one thing to be reducible to another? Roughly speaking, to reduce one domain to
another is to say that the first domain is really nothing over and above the second.
However, there are still at least two things we might mean when we say this.

The first thing we might mean is called explanatory reduction. This holds that
everything in the first domain can be explained exhaustively in terms of the second.
This kind of thing is also sometimes call analytic reduction, in the sense that I can
give a complete analysis of all the phenomena in the first domain in terms of the
second.

Let’s have an example. Take vision. I might think that everything about vision can
be explained or analysed using purely physical terms, such as those relating to the
wavelengths of visible light, for instance, or to the anatomy of the optic system, and
so on. To say that this kind of explanation would be exhaustive is too say that vision
is explanatorily reducible to the physical domain. It doesn’t deny that there is such a
thing as vision, it’s just that an explanation or analysis of that phenomenon given in
purely physical terms wouldn’t leave anything out – we wouldn’t be missing anything
if we were to ‘reduce’ all talk of vision to talk of light waves etc.

The second kind of reduction we might mean is called ontological reduction. When
we say that one kind of thing is ontologically reducible to a second, what we are
saying is that the second kind of thing is the only kind of thing that actually exists
when we really get down to the bottom of things. The first kind of thing, if it is
ontologically reducible to a second, is actually just composed of the (truly existing)
second kind.
I might, for example, think that tables are ontologically reducible to their atomic structures. Although it looks like there are such things as tables in the world, the only things that really exist are molecules, and the thing that I thought was a table is really just a table-shaped molecular arrangement. We can still talk about tables, since doing so seems to be helpful for our conversational purposes sometimes – it would be rather a nuisance if our philosophical metaphysical theories forced us to only ever talk about things in terms of their atomic structure! It’s just that strictly speaking, what we’re really talking about when we have a conversation about tables are collections of atoms arranged in a table-shaped way. Tables themselves don’t, strictly speaking, exist.

There are actually a few more ways of understanding reduction than these two, but these are the only ones that will be needed for our purposes. (See further resources).

**Reduction and supervenience**

Notice that saying that something is supervenient on another is saying something very weak. All that’s needed to make a claim like this true is a systematic dependence of one set of properties on another. To claim something is reducible to another, by contrast, is to make a much stronger claim.

The stronger claim actually brings with it the weaker claim for free. Take the last example I gave. If the table is ontologically reducible to its molecular components, then its table-properties must also supervene on its molecular properties. This is because if tables are really nothing over and above the molecules that make them up, then there can be no difference to whether or not we see something as a table without there also being a difference in how the molecules are arranged. Likewise, so long as there are two exactly similar arrangements of molecules, those molecule-groupings will stand and fall together with respect to whether or not they count as being arranged table-wise or not.

Reduction claims, then, always entail supervenience claims. Given their weakness, however, it’s not really surprising that supervenience claims don’t likewise entail reduction claims. Suppose, for instance, that I think that dancing-properties supervene on physical ones – there can be no difference as to whether and how someone is performing a dance without a corresponding difference to the physical properties of the dancer’s bodily movements. I might assent to this, but still resist the stronger claim that dancing-properties are reducible to physical ones. I might still think that a complete physical analysis of what’s happening will still leave something out, either explanatorily or ontologically speaking.

**Reductive materialist accounts**

Reductive materialist accounts of the mind come in both flavours of reduction – they sometimes say that the mind is explanatorily reducible to the physical, and sometimes that the reduction is ontological. In what follows I will present two different reductive accounts: logical behaviourism and identity theories. These are both reductive accounts of the strongest kind – they are ontologically reductive accounts.
Other than whether they are explanatorily or ontologically reductive, the main difference between different reductive physicalist accounts of the mind lies in the choice what kind of physical property the mind is taken to be reducible to. For logical behaviourists, for instance, mental properties are reducible to physical behavioural properties, whereas for identity theorists they are reducible to physical properties of the brain.

Teaching tips

- A fun way to test the students’ intuitions about whether or not they are materialists is to use the question proposed by Saul Kripke: If God came along and fixed all of the physical facts about the world, would he have more work to do in order for there to be mental properties in the world? Those who don’t think there’s any more work to do are naturally inclined reductive materialists.
- It’s good to have a few examples of properties that might be taken to supervene on the physical to get the students to really grasp this relation. A couple of suggestions for supervenient properties might be biological, chemical, cultural, aesthetic, tactile, communicative, cartographical, representational (almost anything you can think of that isn’t purely physical!). You could try to get the students to think of a few themselves.
- To get a better hold of what the mind-body supervenience claim amounts to, you could use clones/teletransporter/doppelganger cases etc. to cook up scenarios where there is someone exactly physically identical to you in every way. If the mental supervenes on the physical then this entails that your clone will also be mentally identical to you.

Discussion questions

- Can you think of any other properties not already mentioned that supervene on physical properties?
- Can you think of examples where one thing is explanatorily reducible to another? What about where something is ontologically reducible?
- What would a reductive physicalist account of dancing say? (Where it’s explanatorily reductive? And what about where it’s ontologically reductive?)

Further reading

- Stanford encyclopedia has a very detailed, thought slightly technical, entry on supervenience: [http://plato.stanford.edu/entries/supervenience/](http://plato.stanford.edu/entries/supervenience/) that would certainly be useful here.
- The early sections of David Papineau’s chapter on ‘Supervenience and Identity’ also gives a helpful account of what a physicalist means by saying that the mind supervenes on the body. An online copy is available here: [http://www.kcl.ac.uk/ip/davidpapineau/Staff/Papineau/PhilNat2nded/ch1%20revised.htm](http://www.kcl.ac.uk/ip/davidpapineau/Staff/Papineau/PhilNat2nded/ch1%20revised.htm)
- For technical terms like 'supervenience' and 'reduction' it is important to ensure that your understanding of the terms are as precise as possible. Philosophical dictionaries and encyclopedias (other than the Stanford one) that are available online include:
- Oxford Reference Online
- Routledge Encyclopedia (requires login)
  [http://www.rep.routledge.com/LOGIN?sessionid=cabfbb1fa320dc49d30b91cf84d4517a&authstatuscode=400](http://www.rep.routledge.com/LOGIN?sessionid=cabfbb1fa320dc49d30b91cf84d4517a&authstatuscode=400)
- Internet Encyclopedia of Philosophy
  [http://www.iep.utm.edu](http://www.iep.utm.edu)
- Here’s a link to a website that lists some different kinds of reduction in philosophy, that go beyond the two discussed here: [http://www.liv.ac.uk/~pcknox/teaching/phil/phil/sred.htm](http://www.liv.ac.uk/~pcknox/teaching/phil/phil/sred.htm)
- Almost all philosophy of mind introductory textbooks will have a section on reductive physicalist accounts. See, for instance, the section on reductive physicalism in William Jaworski’s *Philosophy of Mind*, Oxford: OUP (2011).
2.1 Logical Behaviourism

The first reductive account I will discuss is called *logical behaviourism* (also sometimes known as *philosophical behaviourism*, to distinguish it from *methodological behaviourism* in psychology). According to logical behaviourism, the physical properties to which mental properties are properly reducible to are behavioural ones. On this view, to ascribe someone a mental state is to ascribe to them a set of behavioural dispositions. To say that someone is in pain, for instance, is just to say that they are currently disposed to display certain characteristic ‘pain’ behaviours – writhing around on the floor, say, or clutching the injured body part and saying “ouch”. To say that someone thinks it's going to be sunny today is to say that that are disposed to wear a t-shirt and pick up their sunglasses on their way out of the door. To say that someone believes the train leaves at 11am is to say that they are disposed to plan their morning accordingly. More generally speaking, we can say that to be in a particular mental state is to be disposed to display a given range of associated behaviours in response to a given range of stimuli.

Saying that these dispositions are associated with these states of mind is not, of course, to say anything particularly revelatory. The crucial point about logical behaviourism is that it says something much stronger than this – it's not just that if you are in pain then you will be disposed to writhe etc.; according to the behaviourist your pain is your disposition to exhibit those behaviours. Using the locution introduced in the last section, we can say that your pain-state is ontologically reducible to your behavioural dispositions. Metaphysically speaking, it is really nothing over and above them. Put this way, behaviourism looks like a pretty striking position, and it might not come as much of a surprise that it is not a particularly favoured one in contemporary philosophy of mind. Saying that, it enjoyed a brief burst of popularity around the middle of the last century, so it must have something going for it. What considerations, then, could there be in its favour?

**Arguments for behaviourism**

**Verificationism**

In a way, the reason for including behaviourism in the standard catalogue of responses to the mind-body problem is really for historical interest. I say this because the main source of its popularity was another philosophical position that was floating around at the start of the twentieth century, that has since been almost universally discredited. This was a theory of meaning called *verificationism*, according to which the meaning of any sentence is given by its mode of verification. The meaning of the sentence, “There are two butterflies on the table”, according to this view, is something like ‘if one was to look at the table one would see two butterflies; if one was to put one’s hand out to the table, one would feel two butterflies; if one listened closely enough... etc.’.
A historical explanation for the emergence of behaviourism as a theory of mind is that it provides a natural companion to this theory of meaning that was so popular at the time. After all, if the meaning of a sentence is given by its mode of verification, then what should we make of sentences attributing mental states to others? What mode of verification can we apply to those sentences? Not introspection – that is only available first personally, so can't be the way to verify our mental state ascriptions to other people. Well, it seems like we normally rely on behaviour to tell us about others' states of mind. According to verificationism, to take this seriously as the relevant mode of verification is to say that the meaning of mental-state attributions is given by statements about behaviour, or behavioural dispositions.

Given that verificationism is no longer accepted, of course, this no longer gives us a reason to espouse a behaviourist view of the mind. With this historical background in place, however, I hope it is more understandable why it was popular for a time. This is not so much an argument for behaviourism as an account of why some people might have gone in for it.

**Explanatory power**

There is another argument that might still be taken to work in its favour (though admittedly an argument so lightweight that it has little hope of standing up to the battery of arguments we are about to see against the view). This is a consideration of explanatory power: the behaviourist, we might think, can explain things that are left mysterious on some other theories of mind – namely, the connection between the mind and behaviour.

It is, I think, pretty intuitive that there is a deep and systematic relation between mental states and their associated behavioural profile, or their characteristic dispositional inputs and outputs. There seems to be something significant and non-arbitrary, for instance, about the link between the state of fear and threatening kinds of stimuli on the one hand, and trembling/sweating/fleeing/etc. behaviours on the other. The behaviourist can easily accommodate this intuition. The reason that we think that there are important connections between mental states and their behavioural dispositional profiles is because those states are nothing over and above those profiles. At the very least, then, we can say that the behaviourist wields some attractive explanatory leverage.

**Arguments against behaviourism**

**Qualia**

The main problem with behaviourism is not difficult to spot: it misses out what it feels like to be in pain. To say that the mind is reducible to behavioural dispositions leaves out what we surely think is one of the most central and ineliminable features of the mental realm – namely, that having a mind, or being in certain mental states, comes with a certain experiential quality; in the famous phrase from Thomas Nagel, *there is something it is like* to be minded (see further
resources). Such experiential properties of the mind – the ‘what-it-is-like-ness’ of a given mental state – are sometimes called **qualia** (a single one of these properties is called a ** quale**).

The first objection to behaviourism, then, is that it cannot account for the existence of qualia. It can’t explain why being in a state of pain feels like something. You might say that just as above there was an argument for behaviourism on the grounds that it has some kind of explanatory power, this is an argument against behaviourism on the grounds that it **lacks** explanatory power with respect to the existence of qualia.

**Superspartans and pretenders**

A second objection to behaviourism is that having the relevant behavioural dispositions seems to be neither necessary nor sufficient for being in a particular state of mind; a crippling objection if ever there was one! Both of these points can be made through thought experiments.

First, why isn't it sufficient? Well, imagine a person who – as an unexplained anomaly – is born without a sense of fear. When confronted with what we would think of as threatening situations, she feels nothing; she is entirely unfamiliar with what it feels like to be afraid. Having grown up in a society of ordinary human-beings, however, she has easily picked up the relevant profile of inputs and outputs typically associated with that state of mind: she has developed a good grasp of the kinds of situations that normally prompt it, and has noted that the usual kinds of responses it provokes include things like trembling, attempts to remove oneself from the situation, darting eye-movements, etc.

In an effort to avoid revealing her birth defect, she perfects a ‘fear’ act in which she accurately mimics these fear-behaviours whenever she recognises that a fear-stimulus is present. The act, let us suppose, is faultless. It is true to say of her, then, that she has a full ordinary set of fear dispositions, which are activated in typical fear situations. Surely, however, we would not want to say that she ever really undergoes the state of fear. The whole point of the story is that she is only pretending. To say that someone has the relevant dispositions, then, is *not sufficient* for saying that they have the mental state in question.

A converse case can be constructed to show that it is not necessary either. In a famous thought experiment due to Hillary Putnam (see further resources), there is a society of ‘superspartans’ who have an unbreakably strong social norm that one must never reveal oneself to be in pain. No matter what agony they might be feeling, these superspartans are trained to remain cheery and impassive. Again, suppose that this learned behaviour is exceptionless. Superspartans, according to this scenario, have none of the typical behavioural dispositions associated with the state of pain. Do we want to say that this means they don’t *have* any pains? Surely not; they are sometimes, as the story has been given, in agony! What this case shows is that we don’t have to say that someone has the relevant kinds of behavioural dispositions in order to say that they’re in pain. Behavioural dispositions *not necessary* for a pain-ascription.
Mental causation

These two arguments against behaviourism are normally taken to be fairly decisive. I will, however, just mention one more: the argument from mental causation. In the arguments against substance dualism, we saw that we have a seemingly unyielding intuition that our mental states and events sometimes cause physical behaviour – we saw my desire for coffee cause my reaching for the coffee cup. If behaviourism is correct, however, then this intuition must be given up on this view too. My desire for coffee didn’t cause my coffee-cup-reaching on the behaviourist’s account, since on that account the coffee-desire was the reaching. Just as the intuition of mental causation caused problems for the substance dualist, then, it likewise causes problems for the behaviourist physicalist.

Teaching tips

- This theory of mind is somewhat removed from our normal understanding of the mind and mental states, as a result of which students might feel some resistance to it.
- Because of this, it’s perhaps worth starting with some mental states for which the view is a little more plausible. It is, for instance, somewhat plausible that mental characteristics like shyness, or being judgmental, or bearing grudges, are plausibly kinds of things that really are just a matter of having the right kinds of behavioural dispositions.
- Likewise, a pretty convincing case could be made for propositional attitudes like beliefs just being the same as having the disposition to behave in certain ways. My belief that the coffee cup is on the table is plausibly just my disposition to reach in that direction whenever I want to pick it up. If students resist, try to get them to articulate what they think is missing from this account. The good thing about propositional attitudes for this view is that there’s plausibly nothing it is really like (no qualia) associated with these states. (Other propositional states you could use include: beliefs, hopes, and regrets).
- The really problematic states for this view are going to be the ones that are strongly associated with certain kinds of experience – primarily, emotions and all forms of sensation.
- To sort through these different cases, you could offer the students a hybrid view on which some mental states should be understood as behavioural and not others, and then get them to brainstorm which should go in which category.

Discussion questions

- Do you think there’s anything right about the behaviourist account of the mind? How would you defend the theory if you had to stick up for it on the behaviourist’s behalf?
- Are there some mental phenomena that are more conducive to a behaviourist analysis than others?
• Taking a mental state other than pain, could you construct a thought experiment that shows that having the relevant behavioural dispositions are neither necessary nor sufficient for having the mental state?
• Which of the arguments against behaviourism do you find most persuasive and why?

Further reading

• For more on qualia, or the what-it-is-like-ness of mentality, see Thomas Nagel’s ‘What is it like to be a bat?’ in The Philosophical Review, Vol. 83, No. 4 (1974), pp. 435-450. Also a good paper to read to get a bit of background to reductive physicalism more generally. (http://cutonthebiasworkshop.files.wordpress.com/2011/05/nagel-1974-what-is-it-like-to-be-a-bat.pdf)
• There will be a chapter on behaviourism in most introductory books on the philosophy of mind. See, for instance, Chapter 2 of Jaegwon Kim’s Philosophy of Mind, Cambridge MA: Westview Press (2006).
2.2 Identity theories

While logical behaviourism reached its peak of popularity in the first half of the last century, identity theory (sometimes also known as central state materialism) was introduced by a small group of Australian philosophers at the university of Adelaide in the late 40s and 50s, and expanded to international popularity among the philosophical community during the 50s and 60s.

In the introduction to non-reductive physicalism we noted that the difference between the various reductive physicalist theories of mind can be drawn in terms of a difference in what kind of physical thing the mind is thought to be reducible to. The answer that the identity theorist gives here is that mental states are reducible to states of the brain.

In a slogan the identity theory is the view that mental states are brain states.

Type/token distinction

Before going any further, it will first be important to get clear on a certain distinction between what we call types and what we call tokens of a given thing. Consider the following inscription:

glee

How many letters are there here? Well, it depends on how you’re counting them. One correct answer would be that there are four letters here – a ‘g’, an ‘l’, an ‘e’ and another ‘e’. If this is the answer you’re after, then it looks like the question you were asking is how many actual letter items there are here. In terms of the type/token distinction, we would say that you’re counting the letter tokens.

This, however, is not the only way you could answer the question. You could also say that there are only three letters here – ‘g’, ‘l’ and ‘e’. If this is your answer, then you are probably interested in how many different kinds of letters there are. We would say that you are counting letter types.

The same kind of distinction can be made when it comes to talking about mental states. Take, for instance, the mental state of thinking that the moon is made of cheese. How many mental states are there here? Well, again, it depends on what you’re counting. If you’re counting mental state types, then there is just one: the state of thinking that the moon is made of cheese. On the other hand, you might be asking something quite different – you might be interested in how many actual events there are, or have been in the history of the universe, in which the thought that the moon is made of cheese was actually had. To get an answer to this question would require some pretty extreme research methods! In principle, however, I hope the idea is clear. We might be counting thought-types (of which there is only one), or we might be counting thought-tokens (which may well run into the billions).
Likewise for brain states. Consider a brain state (let’s call it ‘B’ for short) that involves a specified pattern of neurons firing in the right parietal lobe. Supposing that B sometimes does actually take place in normal human beings, how many brain states should we say there are? Again, your answer to this question is going to depend on what you’re really asking. If you’re asking how many types of B-states there are, then there is just one. If you’re asking how many tokens of B-states there are in the world at any given time, then there may well be many.

**Type identity theory**

Using this distinction, we are now in a position to say that the original identity theory of mind, as developed by J.J. Smart, Place and others in Australia in the 1940s and 50s, was a type identity theory. Take any mental state type you please – thinking of one’s grandmother, experiencing tingling in one’s fingertips, listening to an earworm of Beethoven’s fifth, there is literally an infinite number to choose from. According to the type identity theorist, that mental state type is identical to a particular brain state type. Every time anyone thinks of their grandma, they are activating exactly the same kind of brain state.

The strength of this claim needs underlining. Type identity theorists are not just saying that whenever anyone thinks of their grandma, this mental state causes exactly the same kind of brain state to fire up every time, or that the activation of that sort of brain state will systematically correspond to grandma type thoughts. The claim is that the grandma-thought-type is the brain-state-type. This is why they are called ‘identity’ theories; on their view the mental state type is identical to the brain state type.

**Scientific hypothesis**

The type identity theory is not supposed to fall out of our intuitions about what the mind is in the same way that, for example, we might think that substance dualism does. When it was proposed, it was not done so in the spirit of common sense. It was, rather, put forward as a substantive scientific hypothesis, or as an empirical discovery about the world. To see what I mean, think about the first time that anyone proposed that water is H₂O, or that heat is molecular energy. These are type identity claims, just like the type identity theorist’s claim about mental states and brain states. The first, for instance, says that a certain type of natural stuff in the world – water – is type-identical to a certain kind of chemical composition – hydrogen and oxygen molecules, bonded together in a specific way. This suggestion must have come as something of a surprise to most people when scientists first proposed it, but it didn’t really matter that the claim was not obvious from our intuitions about water. It wasn’t supposed to be intuitive. It was a scientific discovery.

According to the type identity theorist about the mind we should think about the mind-brain identity in the same kind of way. Given the seeming untenability of substance dualism (that emerges from the problems with that view that we’ve already seen), the type identity theorist thinks that we should be open to the
scientific discovery about the mind and the brain: just like water turned out to be identical to H₂O, our mental states turn out to be identical to our brain states.

One important upshot of this status of the type identity theory as a scientific hypothesis is that it means that a certain line of attack against the theory can’t even get off the ground. The line of attack I have in mind says that the mind can’t be identical with the brain, because otherwise we would have noticed this fact. Or, put another way, if mental states are really identical with brain states, how come we can easily pick out different mental states, and yet be completely ignorant as to which brain states they are? I hope it is clear that these kinds of objection are non-starters. Given that the mind-brain identity is a scientific discovery – just like the water-H₂O identity – it is no surprise that a bit of scientific investigation is required before we can work out which mental state type is identical to which brain state type.

Even if this isn’t a very good kind of objection to type identity theory, however, there are a number of much better ones available. Here are two.

**Problems for the type identity theory**

**Leibniz’s law**

In the section on the indivisibility argument, we came across ‘Leibniz’s law’, or the principle that two things are identical if and only if they share all of their properties. The idea behind this principle, remember, was that if two things really are one and the same thing, then how could they possibly differ in the properties they have? That would be the same as saying that one and the same thing both does and does not have a given property!

We can use Leibniz’s law to generate problems for type identity theories; all we have to do is come up with properties that are not shared between the mind and the brain. There are two ways of doing this: we can think of properties that the mind has but the brain lacks, or we can look for properties of the brain that are not had by the mind. At least on the face of it, there seem to be plenty examples of both (but see discussion questions for possible responses). This, really, is very similar to the arguments seen earlier from Descartes for substance dualism.

As an example of the first kind of property, take a pain that I’m feeling in my left foot. That pain has the property of being located in my foot. Clearly, however, this isn’t a property shared with my foot-pain brain state which, presumably, is being instantiated in my brain. We can also go back to those ‘qualia’ again (the experiential properties of the mind). There is something it is like to have a foot-pain, but as far as I know, there’s nothing in particular it’s like to be in brain state B.

Going in the other direction, there are also properties that the brain seems to have, but that the mind lacks. Brain states, for instance, have certain frequencies, but would we really want to say that this is also true of my mental foot-pain?
The type identity theorist may well have ways of defending herself against these kinds of apparent problems, and they are not normally taken to be decisive against identity theories. At the very least, however, there is a challenge here that calls for a response.

**Multiple realizability**

By contrast, *the problem of multiple realizability* is often taken to be fatal for the type identity theory (first put forward by Hillary Putnam, see further resources). One way of putting this objection is to say that it’s really rather chauvinistic of us humans to assume that the mental state of, say, pain is type-identical with brain state B, where brain state B is the kind of brain state that we humans are in whenever we are in pain. Surely, after all, there are lots of animals sensitive to pain whose brains don’t look anything like ours. Would we really want to use this as a reason to deny that they feel pain? Take, for instance, the squid. Squids’ brains are no doubt very different to ours – made from different stuff, differently structured, etc. They couldn’t, then, ever be in brain state B; they just don’t have the right kinds of brains for it. But if type identity theory is correct, then brain state B *just is* the mental state of pain. If squids can never be in brain state B, then the type identity theorist will have to say that they can never be in pain. Poor squid! Given that we don’t really want to say this, then the fact that it comes out as a result of the type identity theory means that we must have gone wrong somewhere along the way.

This objection packs an even more forceful punch once we leave the boundaries of the actual world and look at how things could have been in other possible worlds (which, remember, is a way that we have of testing *metaphysical* possibilities). Imagine that there’s a world whose inhabitants are really similar to us. So similar, in fact, that you wouldn’t be able to tell them and us apart from the outside – they look like us, they have full and tiring lives like us, they fall in love, have favourite breakfast places and so on. Open up one of their skulls, however, and the difference is easy to spot. Where our brains are normally located, these creatures have a sophisticated arrangement of silicone chips. Like the squid, this means that these creatures could never be in brain state B (indeed, they could never be in *any* brain state – they have no brains!). But if brain state B is identical to the mental state of pain, then we will have to say that they can never be in pain either. But surely it’s even less plausible to say this in the case of these creatures, who look and act just like us, than it was with the squid.

This objection is called *the problem of multiple realizability* because it amounts to the claim that it is very plausible that the same mental state (say, pain) could be realized in a variety of physical states (say, human brain state B, squid-state-C, silicone-state-D, etc.). If this is right, then the type identity theorist has a serious problem: given the core commitment of that view of type identity between mental and brain states, the type identity theorist can’t accept such multiple realizability of mental states.
A solution?: Token identity theory

There is a way of refining the identity theory of mind to make it compatible with this multiple realizability of mental states that has been taken by some philosophers intent on salvaging something from the theory. Rather than looking at *types* of mental state, the token identity theorist is interested in mental state *tokens*. Take a specific mental state that you might be currently undergoing – say, a feeling of strong elation. When you undergo that very feeling, you are in a token mental state that can never be repeated. This is so even if you feel the same kind of elation every morning when you wake up – in that case, it would be true to say that you have recurring mental states of the same *type* every morning (namely, the elation-type). No two instances of that type, however, can count as the same token; we would just say that they are both tokens of the same type.

Likewise for brain states. The token identity theorist – unlike the type identity theorist – has no special use for types of brain state. The token identity theorist, rather, is interested in actual neural events going on on a given occasion. They are, that is to say, only interested in brain state *tokens*.

The revised theory says that any given mental state token (e.g. my actual feeling of elation at 19:32 on 23rd June 2013) will be identical with a token brain state (e.g. firing of synapses in position,... in the brain at ... frequency at 19:32 on 23rd June 2013.) Mental states are still brain states on this version of the theory, but the identities can only ever be worked out on a case-by-case basis. There is no expectation that all tokens of a given type of mental state will be realized in a brain state of a common type. The advantages of identity theory, then, (namely, it's strongly scientific and anti-dualist spirit) are preserved, but the multiple realizability objection is avoided.

Teaching tips

- You could use some of our contemporary ways of talking to introduce this kind of theory: ‘She’s really brainy’, ‘she’s got a good head on her shoulders’, ‘where’s your head at?’, ‘I wish I had her brains’, ‘I can’t get my head around it’ etc.
- Be sure to use lots of examples to explain the type/token distinction, and maybe try to get them to come up with a few of their own. It's easy to confuse the type/token distinction with (the different) universal/particular distinction, which is the distinction between a general property like ‘redness’ or ‘beauty’ and specific instantiations of it.
- Like with all of the reductive theories, it is important to be clear about just how strong the claim is. In the case of identity theories, the claim isn’t just that mental states (types or tokens) are systematically correlated with brain states (types/tokens). The mental states *ARE* the brain states. It is easy to grant these theories unwarranted plausibility by unwittingly slipping back to the weaker claim.
• If students do find it completely implausible, however, then this shouldn’t be used as a point against it; just remind them that the identity theorist can appeal to fact that this is a scientific hypothesis, not an intuitive claim.

Discussion questions

• In response to the ‘Leibniz’s law’ objection, would it really be so bad to bite the bullet and to accept that mental states and brain states really do share all these properties? After all, before we discovered that water was H2O, there are surely properties of that chemical composition that we would not naturally have ascribed to water. Why not accept that, as it turns out, mental states really do have a certain frequency?

• One way to extend the discussion on multiple realizability is to ask whether we might even want to say that the same mental state can be differently realised in the same species. This would block a response to the multiple realizability objection on the behalf of the type identity theorist that says that type identities are restricted to within a species.
  o Could there even be differences in the realization of mental states within a single person over time? Just think of neurological patients whose single brain functions adapt to be realized in a new part of the brain following trauma to the old realization site.

• Which is the stronger theory out of type- and token-identity theory? (This is to bring out the fact that type-identity theory actually entails token identity theory, but the converse is not true)

Further reading

• A good place to start on this topic is Tim Crane’s Elements of Mind, Oxford: OUP, 2001, which has a chapter both on identity theory and on the problem of multiple realizability.


• Another useful place to look at the multiple realizability objection is Jaegwon Kim’s ‘Multiple realizability and the metaphysics of reduction’ in Philosophy and Phenomenological Research, Vol. 52, No. 1 (Mar., 1992), pp. 1-26

• For a classic formulation of type identity theory, see J.J.C. Smart’s ‘Sensations and Brain Processes’ in Philosophical Review, 68, pp.141-156. Pdf version here:
http://philosophyfaculty.ucsd.edu/faculty/rarneson/Courses/SMARTJACKphil1.pdf
3. Non-reductive materialism

In the next couple of sections we’ll consider some non-reductive materialist, or physicalist, accounts of the mind. To get a grip on non-reductive materialism it will be helpful to reflect back on what it meant to be a reductive materialist. Reductive materialists, we saw earlier, accept two claims, one weak claim and one stronger one (where, we saw, the stronger claim actually entails the weaker one). The weak claim says that mental states supervene on physical states. The stronger one claims that mental states (or talk of them) are reducible to the physical.

With these notions already in hand it’s not difficult to spell out what non-reductive materialism amounts to. Non-reductive materialists agree with reductive materialists about the supervenience claim, but deny the stronger reducibility claim that reductive materialists accept. For the non-reductive materialist, mental states are not reducible to any kind of physical state, they merely supervene on them. And just to remind ourselves: to say that one thing supervenes on another is to say that there can be no change to what properties there are of the first kind without there also being a difference to what properties of the second kind the object has.

**Supervenience without reduction: an example**

Perhaps an example would help here.

While it’s almost always a matter of controversy whether or not a given phenomenon is reducible, most people would probably agree that the category of aesthetic properties are supervenient on, but not reducible to, the physical features of an object. Take the aesthetic property of beauty. This property is supervenient on the physical features of an object; surely there couldn’t be any difference to whether or not something counts as beautiful without there also being a difference to how things are physically with the object. Even if we go this far, however, we still might not think that the property of being beautiful is reducible to the object’s physical properties. Even if we were to produce a detailed and exhaustive inventory of the object’s physical properties, we might still think that something more is needed before we can settle the matter of whether or not the object counts as beautiful.

Non-reductive accounts of the mind think that mental properties are like this too. Suppose that neuroscientists were able to provide a fully comprehensive report about the physical properties of my brain. According to the non-reductive physicalist, reading this report to me wouldn’t amount to telling me about my mental properties. Even though what mental properties I have might depend on the kinds of physical facts about my brain that are listed in that report, they aren’t reducible to those facts.

**Attractions of non-reductive accounts**
By admitting that there really is something special about the mental realm that can’t be reduced to the physical, non-reductive accounts of the mind come across as far more palatable than hard-nosed reductive accounts. By nevertheless insisting that the mental still supervenes on the physical, however, they avoid the difficulties of substance dualism. Non-reductionism, then, gives us an attractive way to hold on to the thought that mental properties are instantiated by, or are causally dependent on, physical states in some way while stopping short of claiming that the two are one and the same thing. These days, most philosophers would probably count themselves as non-reductive physicalists.

**Mere supervenience theories**

The simplest way to have a non-reductive account of the mind is just to state that mental properties supervene on physical ones (or, more precisely, to state that they merely so supervene, given that identity also counts as supervenience and supervenience accounts are designed to stop short of identity claims).

To say this much, however, is not really to say very much at all. There are, after all, lots of different ways that two sets of properties could be related that would count as one of them supervening on the other. The first set might just be causally dependent on the first, or it might be realized or instantiated by it, or constituted by it, or even identical with it (though, to be fair to the supervenience theorist, I did just say that identity was explicitly ruled out). It also isn’t to say anything about how strong the supervenience is supposed to be. Is it just that as a matter of fact, the first set of properties supervenes on the second (what we might call *de facto* supervenience), or is the supervenience a matter of logical necessity (*de jure* supervenience)? To give an account that simply refuses to further specify the relation between the mental and the physical, then, feels a little unsatisfying. We haven’t explained very much at all.

**Beyond mere supervenience: two more non-reductive accounts**

Most non-reductive physicalist theories of the mind, then, attempt to give an account of the mind-body relation that goes beyond a mere supervenience claim. In what follows I will set out two theories that have done this: one that says that mental states are categorisable by their ‘functional’ profile regardless of how it is physically manifested (a view known as *functionalism*), and one that says there is simply no explanation of the mental realm to be given at all that would link it to the physical realm in a systematic or lawlike way (so-called *anomalous monism*).

**Teaching tips**

- Refresh students’ understanding of the notions of supervenience and reducibility with plenty of examples.

**Discussion questions**

- Can you think of any other properties that you think *supervene* on the physical domain, but are *non-reducible* to it?
Further reading

- See the Stanford entry on supervenience: [http://plato.stanford.edu/entries/supervenience/](http://plato.stanford.edu/entries/supervenience/). As with some of the other technical terms, it's worth having a good look at the various philosophical dictionaries and encyclopaedias to make sure you have an accurate understanding of it.

3.1 Functionalism

The first non-reductive physicalist account I will discuss is a view called *functionalism*. This view was first developed in the late 1960s and 70s and I think it’s fair to say that it has never really fallen back out of general favour. The basic idea of functionalism is that by trying to pin down what kind of stuff mental states are made of, both substance dualists and reductive physicalists are going about things the wrong way. Mental states shouldn’t be categorised according to what they’re made out of but, rather, by their causal, or functional, *role* – by what kinds of inputs and outputs it has.

This view is *non-reductive* because its supporters don’t think that mental states are reducible to physical states. Talk of mental states on this view, rather, amounts to talk of the inhabitants of certain functional or causal roles, which themselves might be manifested in any number of ways. A full inventory of the physical facts about the world, then, won’t be enough to tell you all the mental facts too; the latter cannot be reduced to the former (either ontologically or explanatorily).

Functionally definability

Some things in the world offer themselves up to what we’ll call a *functional* characterisation. Let’s introduce a couple of examples to demonstrate what I mean. Take the example of necklaces. What is it to be a necklace? Well, necklaces come in all sorts of shapes and sizes, and can be made out of lots of different materials. You might have a necklace in the form of a delicate gold pendant, or it might be a chunky string of glass beads, or a strand of peals, or even a daisy chain made of flowers. The point here is that what matters when it comes to settling the question of whether or not something counts as a necklace is not what it’s made of. What matters is *whether it does the job that necklaces do*. Does it hang around your neck and look aesthetically pleasing? If it does, then you’ve got yourself a necklace no matter what it’s made of. A necklace, it seems, is a *functionally definable* kind of object; we can define what it is in terms of the role it plays rather than in terms of what it’s made of.

Not all objects are functionally definable in this way. Take, for instance, the question of what it is to be a pearl. Is it alright in this case to say that we don’t really mind what it’s made of? That we will count it as a pearl so long as it performs the pearl-role, or does the job that pearls do? Certainly not. To be a pearl is to be composed of calcium carbonate crystals that have been hardened over time in the soft tissue of a shelled mollusc, and nothing else will do. There are, then, some kinds of objects that are simply unsuitable for functional characterisation.

Mental states as functionally definable

Identity theories – which we’ve already looked at – attempt to categorise mental states like pearls rather than like necklaces. This is to say that they offer a *non-functional* definition of mental states. In the same way that peals simply are
calcium carbonate crystals, for the identity theorist mental states just are brain states (where we might be talking either type or token, depending on what kind of identity theorist you are.)

The functionalist, by contrast, wants to say that characterising mental states is more like characterising necklaces; they are kinds of things that we can pick out by the functional role that they play. Just like something counts as a necklace as long as it fulfills the necklace-role no matter what it’s made out of, it doesn’t matter what something is made of, something counts as a mental state of a particular kind so long as it fulfills the functional role associated with that sort of mental state.

An example

Take the example of pain. What makes something a mental state of pain? According to the functionalist, it’s not the fact that the state is realised in a particular brain state. For the functionalist, rather, what makes something the mental state of pain is that the state plays a certain kind of role. We might, for instance, say that the role of a pain state is to respond to tissue damage with both certain kinds of further mental states (anxiety, self-pity, thoughts of retaliation, etc.) and with certain kinds of outwards bodily behaviours, like saying ‘ow’, crying, clutching the injured body part, and so on. As long as this kind of causal profile is instantiated, then I will count as being in pain, regardless of how that state is physically realised – be it brain tissue, silicone chips, squid-brain, or anything else.

Notice that as long as the functionalist thinks that these states are physically realised somehow, then they will have to ascribe to some kind of token identity theory. Each occurrence of a pain state – identified by the functional role just described – will as a matter of fact be identical to a particular token physical state. The difference with straight token-identity theorists, however, is that functionalists don’t think that this identity plays any role in what makes the state a pain state. All of that work is done by the state’s functional role. It just also turns out that the state so characterised will have to be physically realised somehow.

What is the difference with behaviourism?

This, then, is the difference between functionalism and token-identity theory, and I hope that the difference with type-identity theory is pretty straightforward. It might not, however, be so obvious how it differs from behaviourism. Don’t they both identify mental states with the kinds of behaviours they cause?

As it happens, there are at least two important differences between them.

The first difference is like the difference between the necklace and the pearl. For the behaviourist, our mental states are like pearls. On their account, my mental state of pain just is my writhing on the floor, holding my foot, etc. This kind of
behaviour, or my disposition to engage in it, is what my state of pain is, just as a pearl just is a calcium carbonate deposit.

Functionalism is not like this, because functionalism doesn’t say anything about what my mental state is made up of. My state of pain, for the functionalist, is more like the necklace – what counts as my state of pain is just whatever it is that performs the right kind of function, just as what counts as a necklace is just whether or not it plays the right role.

The second difference is in the kinds of ‘inputs’ and ‘outputs’ that are permitted into the analysis. For a behaviourist, these inputs and outputs will all be outwardly perceivable stimuli and behaviours – the observability of these behavioural profiles, remember, was one of the advantages of behaviourism. Such restrictions, by contrast, don’t hold for the functionalist, who can also list other inner mental states and events as part of a state’s functional role.

**Physicalist or dualist?**

By assimilating their account of mental states to the ‘necklace’ kind of explanation rather than to the ‘pearl’ one, functionalism seems be a different sort of account to the others we’ve seen. To see this, notice that functionalism is – at least in principle – consistent with both physicalism and dualism. To say that the mental state of pain is identical to whatever it is that performs the pain-function isn’t yet to settle the question as to whether that thing is going to be a physical substance or a non-physical mental one. If it turns out that the thing that performs that role is a ghostly Cartesian substance, then a dualist functional theory would be the right one. If, on the other hand, it turns out to be a physical thing, then we should be physicalist functionalists.

As far as I know, there aren’t actually any substance dualist functionalists out there. The point here is only that strictly speaking, functionalism doesn’t entail physicalism, so if you wanted you could be both a dualist and a functionalist. The reason that – as a matter of fact – all functionalists have been physicalists, then, must come from the independent reasons to favour physicalism over substance dualism that we’ve already come across.

If, however, you are a physicalist and a functionalist, then you will have earned yourself an extra argument against the substance dualist, or, at least, you will have a way of responding to the conceivability argument that might not be available to other physicalists. That argument, remember, was based on the conceivability of having a mind while lacking a body. Given that conceivability is supposed to be a good guide to metaphysical possibility, this was supposed to show us that a separation of mind and body is metaphysically possible. But identities are metaphysical necessities. So the mind is distinct from the body. Functionalism gives us a way to respond to the first step of this argument – namely, to the claim that we can imagine undergoing a mental state in the absence of a body. If the mental state is, as the functionalist would have it, whatever performs the relevant function, then the functionalist can just respond
that this isn't really a case of imagining a non-physically instantiated mind. Rather, there must be some other kind of physical stuff around doing the job.

**Advantages of functionalism**

Functionalism's biggest selling point is that it accommodates the *multiple realizability* intuition that caused such problems for type identity theories. Just as a quick refresh, recall that this was the idea that a single given type of mental state – say, the state of pain – will plausibly be realizable in different species (and possibly even in the same species) in a whole range of different kinds of physical state. If the intuition that this is possible is right, then type identity theory is in trouble: you can't hold onto multiple realizability while at the same time saying that mental state types are identical to types of physical state.

The functional theory of mind is perfectly designed to accommodate the multiple realizability of mental states. Since for the functionalist pain is whatever performs the pain-role, there is plenty of scope for this turning out to be, say, brain-state-B in humans, silicone-state-C in otherworldly creatures, squid-state-D in squids, and so on. Multiple realizability intuitions, then, should lead us to welcome a view like functionalism.

**Problems for functionalism**

Functionalism says that mental state is identical to whatever performs the right kind of functional role. Its flexibility with regards to what this thing is what makes the view so appealing in the face of the multiple realizability intuition – there are lots of different things that it could be, so we don't have to chauvinistically restrict these mental states to creatures with brains just like ours.

This virtue of the account, however, also ends up being its vice: it is this very liberalism that is the source of the most common kinds of complaints against functionalism. The view, this kind of objection goes, is *too* liberal, and by going in for it we will end up ascribing mental states where we really shouldn't. Here are two examples of this kind of problem for functionalism.

**The Chinese Nation**

This argument, which was most famously given by Ned Block in the late 1970s, is based on the thought that we could arrange systems to perform exactly the same functional role as a given mental state, but where most of us would strongly resist the claim that any mental state is being had.

More specifically, he envisages a system made up entirely of individual human beings, in which a particular mental state role is performed. His is example is often called the ‘China brain’ or ‘the Chinese nation’, because Block uses the inhabitants of China to illustrate his argument. The idea is this. Imagine that every member of the Chinese population is linked up to at least one other member via a two way radio signal, and moreover, that there is an ‘entry point’
into this system that is connected, again via a two way radio signal, to a human body. That body communicates stimuli into the system, and performs output behaviours on the basis of instructions coming out of the system. When a stimulus of, for instance, a ticklish-feeling is introduced into the network, communication between individuals in the system progresses in a way that exactly mirrors the pattern of neural signals that really do fire up in my brain when I actually get tickled. After the requisite activity has passed through the system, suppose that a signal returns to the body from the network instructing it that it should perform behaviours of giggling, flushing, seeking to avoid the source of the tickling, and so on.

Together, the Chinese populants in this story perform the functional role associated with the mental state of feeling ticklish – a ticklish stimulus is fed into the system, and the right kinds of behavioural outputs result. (We might additionally suppose that it will also result in fresh patterns of signaling activity within the network to parallel other mental states caused by the ticklish feeling). But on the functionalist’s account, remember, a mental state is identifiable as whatever it is that performs the appropriate functional role. In this case, this is the population of China! Most think that this is a pretty unpalatable result, which means that the functionalist theory must have gone wrong somewhere.

One of the reasons that we might resist the functionalist’s reading of this scenario might be because we think that there is something important being missed out: namely, that there is something it is like to feel ticklish (those pesky qualia again!). Not only does this thought experiment expose the apparent weakness of functionalism that it is overly liberal in what it will allow in as a mental state, then, it also makes salient the fact that this view seems to have little to say about how we should account for the conscious experiential features of mental states. (See the Shoemaker text in further resources for more on this line of objection).

**The Chinese Room**

A second argument against functionalism is John Searle’s well-known ‘Chinese room’ thought experiment (see further resources). Strictly speaking, this is an argument against what he calls the ‘strong AI’ thesis, or the thesis that appropriately programmed computers can have a mental state such as that of understanding. The argument is, however, equally forceful against the functionalist theory of mind, as we will see.

The thought experiment begins with an imagined scenario that would seem to support the strong AI thesis. Imagine that there is a computer which is programmed in such a way that it can input strings of Chinese of characters, and output sentences of English. If, as functionalism has it, the state of understanding Chinese amounts to nothing more than the instantiation of the relevant ‘understanding Chinese’ functional role, then this computer will count as understanding Chinese.
You might already be loath to accept this functionalist result. Surely, you might protest, it’s a computer, and computers don’t ‘understand’ anything at all. At this point, the argument is a bit like the Chinese nation argument; it works by simply showing that the functionalist view of the mind will allow some systems to have mental states where our intuitions say there aren’t any.

The Chinese room argument, however, has an even more convincing consideration than this up its sleeve against the functionalist view of the mind. This is because it takes the thought experiment a step further in a way that gives us an additional reason to think that this kind of functional role is not sufficient for the mental state in question. What Searle tells us is that he himself (a non-Chinese speaker) could perform this function just as well as the computer. All we would have to do is stick him in an isolated room with the computer’s instruction manual and a load of paper and pencils, and he could translate inputs of Chinese symbols (slipped under the door) into sentences of English. But, as he keeps insisting, there is no question as to whether he understands Chinese – he really doesn’t! What this thought experiment is supposed to show, then, is that performing the relevant function is not sufficient for being in the relevant mental state.

Teaching tips

- Functionalism was at least in part inspired by work in artificial intelligence in the 1950s, and more particularly, by the work of Allan Turing, who invented the ‘Turing test’ for consciousness. This kind of thing might provide an engaging way into the topic.
- With regards to the ‘liberalism’ of functionalism, it is worth trying out different examples on the students to see where they draw their intuitive line between being minded and non-minded. Starting in different directions can urge students in different ways.
  - If you begin with normal computer that has been programmed with functions exactly similar to some of our mental states, for example, then it is likely that most students would not want to ascribe them mental states. Would it change anything to give this computer a robotic ‘body’ that allows it to move around the world, etc.? Probably not. What if we replaced that robotic body with human flesh? Still, why would that make a difference?
  - However, another way of testing this intuition starts in the opposite direction. If I was to replace just one of your neurons with a silicone chip, would that affect whether or not you count as minded? Surely not. But what if I replaced 10 of them, or a thousand, or all of them? Why would that change anything?

This kind of exercise brings out the fact that our intuitions about when something does and does not have a mind are not altogether very reliable.
- Have lots of examples ready to show the consistency of functionalism with dualism and physicalism (angels, God, ghosts, aliens, etc.)
Discussion questions

- What objects can you think of that are functionally definable? And that aren't functionally definable?
- Can you come up with some examples of functional roles for different mental states? (Remember to include both physical behaviours and other mental states in their outputs).
- Can you come up with your own example showing the problematic liberalism of functionalism? i.e. can you come up with an example where there a functional role normally associated with a given mental state is realized, but where we would not want to say that a mental state is being had?
- What could the functionalist say in response to the Chinese nation and to the Chinese room arguments? Are there ways to make palatable the thought that there really are mental states here?

Further reading

- Jaegwon Kim's *Philosophy of Mind*, Cambridge MA: Westview Press (2006) also has a couple of chapters on functionalism, and distinguishes between different types of functionalism, so might be particularly useful for information on these kinds of differences between different functionalist views.
- Stanford encyclopedia has a very detailed article on functionalism - http://plato.stanford.edu/entriesfunctionalism/ - and is another good source for some of the differences between different kinds of functionalism.
- For some key objections to functionalism, see Ned Block's 'Troubles with functionalism' in *Readings in the Philosophy of Psychology, Volumes 1 and 2*, Cambridge, MA: Harvard University Press, pp. 268–305
- See Sydney Shoemaker's 'Functionalism and Qualia' in *Philosophical Studies* 27:291-315 (1975) for an exploration of the objection that functionalism can’t account for states that have associated qualia.
3.2 Anomalous Monism

Another well known non-reductive physicalist account is Donald Davidson's anomalous monism, which can be seen as an attempt to accommodate both the physicalist claim that all events (including mental ones) are fully describable in purely physical terms, and the seemingly conflicting intuition that mental events sometimes cause physical events. (We've already come across this tricky combination of claims in the problem for substance dualism of the causal closure of the physical).

According to Davidson's resolution of this tension, the token-identity theorist is right to say that mental events are token-identical to physical events. This accounts for why physics can fully explain everything that happens in the world. The twist to Davidson's account, however, is that it allows no lawlike explanatory or predictive relations to hold between these physical events and the mental events that they are token-identical to. If we had the time and resources, the idea is, we could give an exhaustive account of all of the events in the world at any given time, including all the mental ones. Nothing we could say using these physical terms, however, could help us to explain or predict what these events would be under their mental aspect. This is because there are simply no laws that could capture any kind of systematic correspondence between certain kinds of physical events and certain kinds of mental ones.

Some terminology: anomalous, monism, non-reduction

Just to be really clear, let's go through some terminology that's already cropped up. First of all, why does Davidson call this position anomalous monism?

The account is monistic, because it takes it that there is only one kind of stuff in the world. This can be contrasted with substance dualism, which is dualistic because it thinks that there are two kinds of substances in the world – the physical and the mental. More specifically, the anomalous monist thinks that the only kind of substance there is in the world is the physical stuff. Because of this it is a physicalist monistic account.

That accounts for the 'monism' part, but what about the 'anomalous' bit? Well, the word anomalous is derived from the Greek word nómos, meaning 'law', or 'custom'. If something is nomological, it is lawlike. If it is anomalous then it is not lawlike. Just think of the word 'anomaly', which we use to describe something that doesn’t fit in with a pattern or lawlike regularity. Davidson’s physicalist monism is anomalous, because it takes it that mental states, events and processes do not correspond to physical events in a lawlike way.

Finally, I said that this is a non-reductive account of the mind. Why is that? Well, Davidson thinks that all mental occurrences supervene on things that happen in the physical world. He has to say this given that – as a physicalist monist – he thinks that physical stuff is the only kind of stuff that exists. He doesn’t, however, think that the mental events will be explained by a complete physical account of the world. In the absence of lawlike connections that could tell us
which kinds of mental events correspond to which physical ones, there is no way of telling which mental states and events there are from the physical facts. While the mental supervenes on the physical, then, it is not reducible to it. And this is just this combination of commitments that we used to characterise non-reductive physicalist accounts.

An inconsistent triad

Davidson argues for his position using what’s known as an inconsistent triad (also sometimes known as a trilemma). An inconsistent triad is a list of three claims, all of which are either well supported or intuitively compelling. The claims, however, are in tension with each other – it looks like they can’t all be true at the same time. This is why we call it an inconsistent triad: it is a puzzle composed of three inconsistent claims, where no single one of them looks like an obvious candidate for elimination.

This is a pretty common way to generate puzzles or problems in philosophy. Normally, the way to solve an inconsistent triad – the way to restore consistency – is to drop one of the claims, and to show why dropping it is not as costly as it might seem. As we will see, however, Davidson does not take up this common strategy. Instead, he presents an imaginative and powerful way to render the three claims of his inconsistent triad consistent.

The three claims

Here are the three claims:

(1) The principle of causal interaction
(2) The principle of the nomological character of causality
(3) The anomalism of the mental

What do each of these mean?

Claim (1) is one we’ve already come across in earlier arguments. It says that there are at least some mental events that sometimes causally interact with the physical domain. We’ve already spent some time urging this intuition elsewhere, but if you want a quick reminder about why we might want to endorse this claim have a look back to the interaction problem for substance dualism.

Claim (2) says that wherever one thing causes another thing, this must happen in a strictly lawlike way. Although this claim has sometimes been questioned, this seems at least on the face of it like a pretty plausible claim. I’ll use an example to try to show why this is.

Suppose that I throw a pebble into my pond in North London, and just at the moment that it hits the water, a storm erupts in Leicester. It would, I take it, be a pretty tall task to persuade anyone that my pebble-throwing caused the storm.
Consider now, however, a development of the case where it turns out that a storm reliably and systematically erupts every time I throw a pebble into my pond. Suppose, that is to say, that we can establish the presence of a lawlike, or nomological connection between my pebble-throwings and the storm-eruptings. Once we add this into the case then it stops looking quite so ludicrous to say my throwing the pebbles is what is causing the storms to erupt, even if we don't have any idea how the causal connection between them works. The principle that causality is always lawlike, then, seems to be one that is supported by our intuitions about when we can say that one thing causes another.

(Notice, of course, that the principle of the nomological character of causality only says that where there is causality, there will be lawlike relations; it doesn't commit to the converse that whenever there is a lawlike connection, there is causality. So the only thing we have to make plausible here is the thought that if there was a causal connection between pebble-throwing and storm-erupting, then we would also expect there to be lawlike connections between them. We don't have to show that the presence of such connections entails that there is causality present).

Claim (3) is the thesis that there are no strict laws on the basis of which mental events can either be predicted or explained, or on the basis of which mental events can themselves be used to predict or explain other events. Another way of putting claim (3) is to say that there are no such things as strict psychophysical laws, or laws connecting the mental realm to the physical. Although Davidson first introduces this claim as an assumption about which, presumably, he thinks we would mostly be in agreement, he does later give us two further arguments for the claim. (Actually, there are more than two arguments for it, but the two main ones will do for our purposes).

The first is the consideration of the holism of the mental. Something is holistic if treating or explaining things about its parts can only be carried out by reference to the whole. (Think of holistic medicine, for instance, that works by focusing on a patient’s whole wellbeing, rather than on individual ailments). The holistic argument for the anomaly claim is that we can only use mental events to predict and explain physical events when we look at our mental lives holistically. My mental event of thinking it’s raining, for instance, will only explain why I pick up my umbrella in the context of a whole range of further mental states (my desire to leave the house, my disliking getting wet, my sense of urgency that prevents me from waiting until it’s stopped raining, etc.). Because of this, we can’t say that the mental state of thinking that it’s raining is systematically, or nomologically, connected to any given physical state. It only leads to any physical events in the world in virtue of its place in a holistic network of mental states.

The second argument comes from the thought that for any given set of physical facts, there are lots of different ways we might interpret someone’s mental states. When we observe someone’s brain-states and behaviour, that is to say, that doesn’t tell us straight-off what’s going on with them mentally speaking. We might call this the indeterminacy of translation between the physical and the mental – even once we have in place a complete list of a person’s physical
properties at a given time, it might be *indeterminate* how we should ‘translate’ that talk of physical properties into mental talk – there might be different mental ascriptions that fit the physical facts equally well. If this is right, however, then there can’t be any strict laws connecting the physical and the mental, since if such laws did exist, then they would give us a straightforward way to settle these kinds of ‘translation’ questions.

**Inconsistency**

Let’s say that we agree with all three of Davidson’s claims. Then it looks like we have a problem. (1) and (2) together entail that mental events must be nomologically connected to the physical, but this is precisely what’s denied by (3). What to do?

**Solution: causal connections vs. nomological connections**

Instead of getting rid of any of these claims, Davidson finds an ingenious way to make them compatible. What we have to do, he says, is just to recognise an important difference between *causal* connections, which apply to things no matter how we describe them, and *lawlike* connections, which only apply to things under certain ways of describing them.

Let’s have an example. Superman and Clark Kent are the same person. Which name we use for that man, however, really depends on how we are thinking of him, or what kind of description we’re thinking of him under. When we call him ‘Superman’, for instance, we are thinking of him under a description like ‘the superhero with the blue and red costume’. This, however, is clearly not how we are thinking of him when we call him ‘Clark Kent’. When we use that name, we are thinking of him as fitting a description like ‘the geeky Daily Planet reporter’.

What does it mean to say that causal connections don’t depend on how an object is described? Well, it means that it doesn’t really matter *which* description we’re using when we talk about things that causally happen to the man we’ve been talking about. If Lois Lane causes Clark Kent to blush, then this is no different to saying that she’s causing Superman to blush; they’re one and the same man, so one of them can’t blush without the other doing so too. Likewise, if Superman causes the villain to flee the city, then so does Clark Kent (even if most people don’t know it). They are just one and the same person, so one of them can’t cause villain-fleeings without the other doing the same. More generally, the moral here is that as long as we have a single object in play, it really doesn’t matter how we describe it when giving an account of a causal event – all that matters is that *that thing* (which we call by both names) caused something to happen.

This, however, is not the same if we are thinking not about causal relations, but about lawlike connections that serve to predict and explain things. Superman’s handsome face, for instance, might serve to explain Lois Lane’s weak knees, or might serve to predict that she will become flushed at the sight of him. For these kinds of connections, it *does* seem to matter that we’ve called him Superman instead of Clark Kent. After all, Clark Kent’s face is, as a matter of fact, the same
thing as Superman's face, but Clark Kent's face can't be used to predict or explain the same things as Superman's face did. Lois Lane doesn't become weak-kneed or flushed when she thinks of the face as Clark Kent's. When it comes to lawlike connections that are used to explain and predict things, then, it does make a difference how we describe them.

**Anomalous Monism**

How does all of this help Davidson to solve the inconsistent triad? The answer comes in two parts.

Firstly, remember that Davidson is a physicalist monist, so he thinks that *all mental events are physical*. This allows him to accommodate claims (1) and (2). Given that mental events are physical events, and physical events are causally connected, we have reason to expect that those physical events that are also mental events will serve to cause, and be caused by, other physical events. Remember that when it comes to causal laws, it doesn’t matter how we describe these events; it will still be right to say that these events cause physical events, whether we talk about them as mental or physical. This is the first part of Davidson's response.

The second part of the answer is that Davidson thinks that it does matter whether we're thinking of these events in their mental or in their physical guise when it comes to the nomological connections that are used to predict or explain things. This is how he manages to also accommodate claim (3), that mental events can't explain or predict, or be explained or predicted by, physical events. This is because explanations are sensitive to the way things are described, and we have already seen that there are reasons to think that when events are described mentally, they can't feature in explanations of, or be explained by, physically-described events. This is the anomalous bit of anomalous monism.

**Token identities**

When I say that Davidson is a physicalist, notice that he can't be anything like a *type* identity physicalist. This is because if certain kinds of mental states were type-identical to certain kinds of physical states, then this fact would serve to provide us with the very kinds of psychophysical nomological connections that he's denying in claim (3). Instead, we can think of Davidson as a more refined and developed kind of token identity theorist. Given his physicalist monism, he will be happy to allow that every mental state is token-identical with a physical state, but given his commitment to the anomaly of the mental, he must resist that there are any type-identities.

**An objection to anomalous monism**

The main objection that has been raised against anomalous monism is the complaint that if you are an anomalous monist, then you can't avoid slipping into *epiphenomenalism*. Epiphenomenalism, remember, is the view that mental states, events and processes are causally inefficacious.
In a way, this is a bit like the problem that we came across in the *causal closure of the physical* objection to substance dualism, which said that despite our strong intuitions to the contrary, mental properties can’t really cause anything in the physical realm, because physical events are always already fully causally accounted for within the physical domain. The objection here is pretty much the same. Given that physical events are fully causally explained by other physical events, what work is there left to do by the mental stuff? They must be causally redundant, so epiphenomenalism must be correct. However, most people think of epiphenomenalism as a result to be avoided at all costs, so would not be happy to accept it as an upshot of their view. Anomalous monism is in trouble, then, if really results in epiphenomenalism.

Davidson does have some things to say in response to this objection which I will not cover here, but see further resources.

**Teaching tips**

- This position is among the trickier ones among the ones covered in these note, but almost everything needed to explain it will have already been covered in the foregoing topics, so this probably one to teach towards the end of a course. (E.g. token-identities, non-reductive physicalism etc.)
- The main new thing that will need to be introduced here is the thought that you can describe things in different ways, and that those descriptions don’t matter to causal questions, but do matter to explanatory or predictive questions. Lots of examples will help here.
- This position offers another opportunity to introduce students to a new form of argument – the inconsistent triad. Have a few simple toy examples of inconsistent triads ready to demonstrate how they work as an argument-type.

**Discussion questions**

- Can you give some examples of subject-areas or topic-areas you learn in your other classes that you might describe as being ‘nomological’? (E.g. in languages there are the rules of grammar, in the sciences you might learn the laws of physics, chemistry or biology, mathematical rules etc.) What about examples of ‘anomalous’ subject- or topic-areas? (E.g. exceptions to grammar rules or vocab learning in languages etc.)
- Do you think that any of the claims making up the inconsistent triad could be rejected? If so, how would you justify its rejection?
- Can you think of a case where one kind of thing causes another kind of thing, but where there aren’t any systematic connections between them? (If not, do you agree with claim 2?)
- Can you come up with a case of your own where one thing causes another thing to happen, but where your explanation depends on the way you describe it?
- Is epiphenomenalism really such a terrible result? Could you reconcile yourself to thinking that your mental properties don’t cause anything?
Further reading


- There is also a clear and detailed entry on it in the Stanford encyclopedia: [http://plato.stanford.edu/entries/anomalous-monism/](http://plato.stanford.edu/entries/anomalous-monism/), which is also a good source of further readings on the topic.

- The key classic reading for anomalous monism is Donald Davidson's 'Mental Events’ in L. Foster & J. W. Swanson (eds.), *Experience and Theory*. Humanities Press (1970), in which Davidson sets out the above argument for anomalous monism. (Online version available here: [http://fitelson.org/proseminar/davidson.pdf](http://fitelson.org/proseminar/davidson.pdf)).

4. Eliminative Materialism

Eliminative materialism, or eliminativism, is really the physicalist theory par excellence: it predicts that some or all of what we currently think of as mental states will eventually simply be dropped from a more developed scientific understanding of the world. Mental states, that is to say, will be eliminated from a mature scientific theory.

This claim is far more radical than any of the physicalist theories of mind that we’ve seen so far. Even the reductive physicalists still thought that there are such things as mental states, or that talking about mental states is still useful for our purposes. For them the point was that when we talk about them, we are really be talking about physical phenomena, or things that can be fully analysed in physical terms. For eliminativists this is still too concessive. They think it’s really only a matter of time and scientific advancement before we realise that the things that we commonly think of as mental states in fact have no place in a scientifically more complete and accurate worldview.

Folk psychology

It is important to start by getting clear on the target for elimination. Roughly speaking, the eliminativist takes as their mark our everyday, commonsensical ways of talking and thinking about our mental lives, or the kinds of states and processes that feature in so-called folk psychology. We might think of folk psychology as a kind of theory that we use to explain and predict behaviour. Suppose that I see Katy running towards the train station, clutching a ticket to Peterborough in her hand. I need not be baffled by the scene – I have a little theory that explains it all. Katy, I might theorise, believes that she’s running late for her train and is clearly very keen not to miss it. She wants to go to Peterborough, and has reasoned that there is a train at this station that will give her a way of doing that. This kind of folk psychological explanation makes ready use of talk of mental states: belief, keenness, wanting, reasoning etc. It is these things that we appeal to in our folk psychological theories, or what we might call the ‘posits’ or ‘postulates’ of those theories, that the eliminativist wants to get rid of.

I should point out that this explanation for how we explain and predict behaviour is not the only one available. It amounts to what’s called the theory-theory – a theory that says that we use a theory to make sense of people’s behaviour. The main competitor to the theory-theory is called simulation theory, which (very roughly put) holds that we predict and explain others’ actions by effectively putting ourselves in their place and asking ourselves what we would do. Given that the eliminativist assumes that we ascribe to theory-theory, one way to refute the view is to reject that assumption. I have included some references for the theory-theory vs. simulation theory debate in the further resources, but this way of arguing against eliminativism involves introducing a fairly substantial new topic area, so this might not be the most profitable line to pursue. Other arguments against eliminativism are noted below.
To recap, then, the eliminativist’s target for elimination are the mental states, events and processes that we postulate in our folk psychological explanations of each others’ behaviour.

**Elimination of mental states**

Eliminativists think that these components of our everyday psychological explanations are ultimately eliminable. Although this is an undeniably radical suggestion, it might not be as outrageous as it first looks. Just think: in the early modern period everyone believed in witches. In a way, the postulation of witches was a bit like our postulation of mental states. It's not that they had hard evidence for their existence, it's just that as long as people could appeal to them, they had an explanation available for some of the otherwise seemingly inexplicable misfortunes that befell them. This is to say that they were theoretical posits, entities featuring in a theory used to explain something. Nowadays in most parts of the world we think that this theory was false. Although we can appreciate why people posited witches – they seemed to explain things that they otherwise had no explanation for, like freak weather ruining crops, etc. – we no longer think they were right to do so. Our scientific theories have developed a lot since then, so we now have much better explanations of these things than the witch-theory gave us.

The eliminativist challenges us to say how we can we so sure that we’re not in the same position when it comes to mental states. Once our scientific theories have sufficiently matured, what's to say that we won’t end up discarding serious talk of mental states in just the same way that we have discarded serious talk of witches? Just as it turned out that there were really no such things as witches, the eliminativist predicts that it’s going to turn out that there were really never any such things as mental states.

**Arguments for eliminativism**

Given that eliminativism amounts to a proposal to discard a particular theory and its posits – namely, the theory of folk psychology and its mental state posits – it is only to be expected that the arguments in favour of it centre on the theoretical failings of that theory. This, however, might come as a bit of a hard sell to students who have not had much experience with weighing up theories on the basis of their theoretical virtues and failings. It can be, for instance, harder to make this kind of argument appealing than the kinds of arguments we saw around substance dualism that relied heavily on our intuitions. The trick, I think, is just to have lots of other examples of theories in hand to try to show students that they already do judge the valuableness of theories on the basis of these kinds of considerations.

**Stagnation**

One of the tests for a good theory in the sciences is whether or not it gives rise to fertile research. A research area is fertile when it leads to a good amount of further active research, whereas a research area is stagnant when no further
progress is being made in it. In a way, this is the objection that the folk psychological theory of mind is a bit of a dead-end; we don’t seem to be getting anything new out of it. By contrast, neuroscientific theories are flourishing – we’re learning more and more all the time about how our brain activities correspond to our behaviours. The eliminativist recommendation is that we should drop the stagnant theory in favour of more fertile ones that ultimately promise us a much more complete understanding of the phenomena.

Support from other theories

Another general theoretical problem with the theory of folk psychology is that it doesn’t seem to be much supported by other adjacent theories. In general, a theory gets much of its support by featuring in a kind of mutually supportive network of theories. The theory, for instance, that dinosaurs became extinct about 66 million years ago because of an asteroid hitting the earth is supported by a number of other nearby theories: natural selection theories, for example, or theories about extinction patterns, and geological theories about asteroid impacts on the earth’s surface and so on. Our dinosaur-extinction theory gets support from consistency with the predictions of neighbouring theories.

Folk psychology doesn’t seem to be like this. What other theories do we have that would support the postulation of mental states? Indeed, the prognosis for folk psychology seems to be even worse than this. For didn’t our discussion of the causal closure of the physical show that our postulation of mental states is not only unsupported by neighbouring theories, but in fact even causes problems for them?

Explanatory power

A third kind of virtue on which competing theories can be judged is their explanatory power. All other things being equal, where one theory explains something more extensively than another, we should choose the first over the second. Take the theory of witches again, which was commonly used to explain the occurrence of disease. When theories about viruses and infections were later developed, one of the reasons why those latter theories won out against the earlier witch theories is because they had greater explanatory power. They could, for instance, explain why a disease will spread through a community in the ways that it does, or why certain symptoms tend to go together, or what factors affect whether or not someone becomes ill – all things that were left mysterious on the witch-theory of disease. To be explanatorily powerful, then, is a major advantage of a theory, and to be explanatory weak is a major failing. (We’ve actually already come across this kind of theoretical virtue in our discussion about logical behaviourism).

A third purported failure of folk psychology theory, then, is that it is not explanatory powerful. There is a whole range of phenomena that are left unexplained by our everyday psychological explanations. According to eliminativist Paul Churchland, these include things like mental illness, creativity, sleep, vision, memory and learning. We might also want to include things like
akratic actions (where we act contrary to our beliefs about the best way to act) and the kinds of unconscious beliefs and desires that we find in Freudian theories of psychoanalysis. These things, the eliminativist thinks, just can’t be explained using our everyday psychological theories, and this explanatory weakness of folk psychology should give us reason to be suspicious of that theory.

**Arguments against eliminativism**

Eliminativism is a radical and controversial position, and I think it’s fair to say that most philosophers would be extremely reluctant to take it overly seriously. The problem, however, is in pinning down exactly what’s wrong with it. After all, the argument against it has got to be something stronger than mere repugnance towards the view – if you feel some resistance to the suggestion that we should stop talking seriously about mental states, just think about how people in the 16th century would have reacted if you had said the same thing to them about witches.

Given that the arguments for eliminativism were based on the theoretical weakness of folk psychology, we might try to argue against eliminativism by defending the credentials of that theory. This has been the tack taken by many opponents of eliminativism. Other arguments against eliminativism involve arguing that the view is internally inconsistent in some way – that it is somehow unintelligible, or self-refuting. I will not cover these arguments here, since they move beyond the considerations of theoretical virtue that I have been focusing on (but see further resources.)

**Theoretical success**

You might not think it from the eliminativist’s onslaught, but folk psychology is actually a very successful theory. If theory-theory is correct, then we use it all the time, in every one of our interactions with each other, to predict what others will say or do. Even in the simple matter of having a conversation with you, for instance, I will normally theorise that you are a willing conversational participant, and so will plan my conversation on the expectation that you are not about the bolt without warning. Indeed, it is plausibly the remarkable reliability and accuracy of these theories that allows us to navigate the interpersonal world so skillfully. When you think about it, it’s pretty amazing that we manage to coordinate our behaviour that springs from inner mental activity into seamlessly executed interpersonal activities. This kind of thing happens both on a small scale – like when we form a plan to meet tomorrow at noon, or when we move a table together – to the big, like sorting ourselves into communities, and arranging large-scale projects together.

This success of folk psychology is so commonsensical that it is at risk of being overlooked. The point of this defence, then, is to bring it back into the limelight. In fact, it is likely that folk psychology will be able to explain and predict things where neuroscience looks pretty hopeless. Could we explain, for instance, why I keep my promise to you to meet at noon tomorrow by appealing only to my
brain states? Probably not. Could we explain that sort of thing using folk psychology? Yes, very easily – we do so all the time. Even if there are some things that folk psychology does not explain then, there are many more things that it does explain.

**Teaching tips**

- To get the students used to the idea of folk psychological explanations, you could show them a number of everyday examples of behaviour, and ask them to give their own folk psychological explanation of it. Once they have done this, get them to extract the mental state terms that they have used. By doing this, they will have identified for themselves the kinds of things that the eliminativist wants to eliminate.
- As mentioned above, a good way to make this topic engaging is to use lots of interesting examples of different theories. It might be a good idea to be prepared with a few historical examples of discredited theories in which some of the entities of the theory have been discarded – stock examples include theories of ether, of phlogiston, demonic forces, etc. Likewise, it might be good to have a few examples of theoretical posits that have not been discredited (e.g. atoms) so that it’s clear that theoretical entities do not always end up being discarded.

**Discussion questions**

- All of these arguments are based on a theory that we’re supposed to apply to others to explain and predict their behaviour. But don’t we have a more direct way of knowing that mental states really do exist? That is, can’t we just introspect them?
  - An eliminativist could respond that we’re so immersed in our folk psychological theories that we unwittingly apply those theories to ourselves too when we take ourselves to be introspecting what’s really there. (Don’t we post-hoc rationalise our actions all the time, and in doing so postulate things like beliefs and desires for ourselves?) The possibility that this is happening makes our introspective evidence pretty unreliable.
- For the general theoretical problems with folk psychology, what help do you think appeal to the soft sciences like psychology, or Freudian analysis might give us?
- Are there more problems for the eliminativist when it comes to states with qualia (experiential properties) like pain, then propositional states like beliefs and desires? After all, we can feel pains, so surely we must be right that they exist?
  - An eliminativist could respond by saying that although it’s right that we sometimes have these experiences, we’re wrong to try to categorise them as ‘pains’. This is because the theoretical postulate of a pain is a confused notion that doesn’t really track anything real in the world, or have a properly worked out explanatory role. People on morphine, for instance, often report having pains, but where they don’t have any unpleasantness
associated with the feeling. Given these kinds of cases, it’s not clear that we’ve hooked on to anything that’s really out there in the world with our notion of a ‘pain’.

- Can you think of any other kinds of mental or behavioural phenomena that can’t be explained by our everyday psychological explanations?
- What could the eliminativist say in response to the theoretical success of folk psychology? Didn’t people in the 16th century also think that the witch-theory was highly successful?
  - Perhaps there are problems with assessing a theoretical framework from inside it. I.e., if we are implicitly assuming the truth of folk psychology then it will be difficult to really assess its success. There is always the risk of ‘confirmation bias’, which is the risk that one will tend to process the data in favour of a working hypothesis by ignoring anomalies, making salient the cases that fit the hypothesis, etc.

Further reading

- Stanford’s entry on eliminative materialism - [http://plato.stanford.edu/entries/materialism-eliminative/](http://plato.stanford.edu/entries/materialism-eliminative/) - both gives a good overview of the view, and also sets out some problems for the view beyond those discussed here.