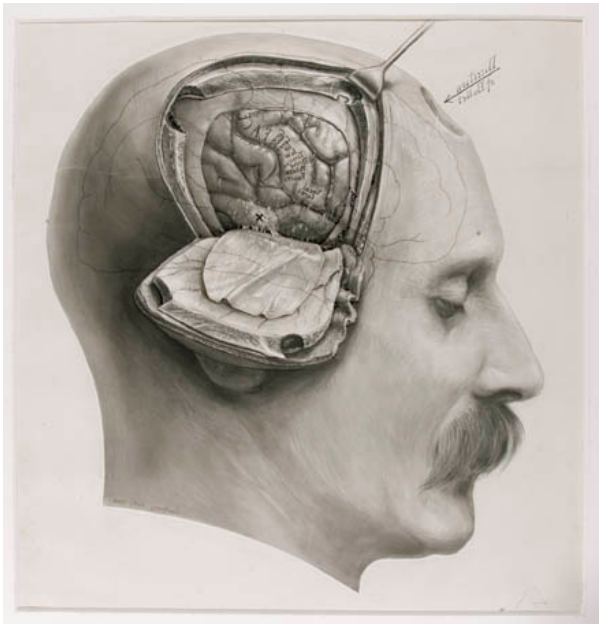
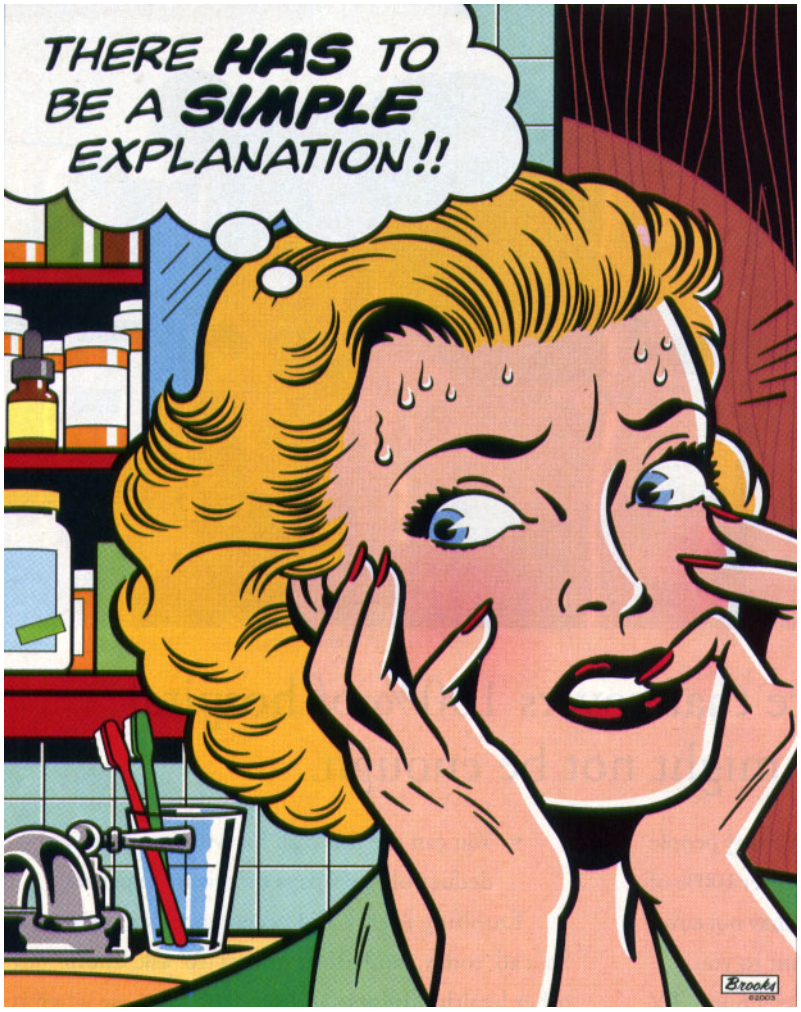


A2 Philosophy of Mind



A note on the final examination

The final A2 examination is very similar in structure to what you sat for AS, although some of the mark allocations are different and (of course) they expect a more sophisticated approach to the material than was required at AS.

As before, there is *one* 3-hour paper, covering both A2 topics (Ethics and Philosophy of Mind). For each of these topics, questions are set as follows:

3 marks: a 'what is X?' or 'what does X mean?' question

2 x 5 marks: questions asking you to 'briefly explain' or 'briefly outline' a philosophical theory or argument.

12 marks: a more detailed explanatory question (one and a half sides of the answer booklet), for example asking about the similarities and differences between two theories, or asking you to explain how a theory can be applied to a particular situation. (Example: how might a utilitarian attempt to justify a ban on fox hunting?)

25 marks: a full essay (at least three sides of writing). This will be an evaluative question about the success or failure of one or another theory you have studied; it could also be a wider 'synoptic' question which gives you a chance to choose which material to talk about, such as 'can mental states be explained in terms of physical states of the brain?'

You'll notice that – with 50 marks per topic – the total mark is out of 100. This also means that your two long essay questions will make up *half* of the total available marks. For that reason you will need to continue to work on your essay style as the course progresses this year.

A note on the handouts

You'll notice that some important words and phrases are in **bold type**. These are important technical terms which you need to be able to explain and/or define as they might be the topic of a 3-mark question in the exam. Where longer sentences or parts of sentences are in bold, this is usually to mark that they are the precise form of words used in the exam board's specification; you should make sure you understand these fully!

Week 1 – introduction to terminology; what is the 'mind-body problem'?

A central tenet of many popular systems of belief, including primitive and classical mythologies, ancient Stoicism, and many currently popular world religions, is that we have a 'soul' which somehow survives the death and decay of our bodies, going on either to some form of afterlife or to be reincarnated in a new body. Such views raise obvious philosophical problems: how could anyone know that this is the case? In what sense could my 'soul' remain the same although reincarnated in someone who has none of my memories?

What is less obvious is that even people who *reject* the existence of immortal 'souls' still face philosophical questions – not about souls, but about *minds*. It seems that the *mind*, at least, is something that everybody presupposes the existence of, for everybody recognizes the legitimacy of using what we might call *mental predicates* such as '...believes that...', '... is in pain', '...is experiencing...'. These predicates seem, taken at face value, to apply to people's *minds* rather than to any physical part of their body (e.g. their brain); we'd be inclined to say that someone who said, for example, 'my brain believes that knowledge is not justified true belief' would be making a kind of linguistic mistake in attributing to a physical object, the brain, something (knowledge) that is properly attributed to the mind itself.

Given that minds exist, a range of philosophical puzzles spring up: what are minds, and what are they made of? How do minds interact with the physical material of our brains and *vice versa*? If minds are in some way made up of the physical material of our brains, or if the existence of our mind is the result of the physical arrangement of our brains, how is this possible? The question of what the relationship *is* between the mind and brain (or between the mental and the physical in general) is known as the '**mind-body problem**', while the question of how conscious mental experience results from the operation of physical processes within our nervous systems is the '**problem of consciousness**'.

Of course, some people (both neuroscientists and some philosophers) have attempted to avoid facing up to these problems by denying the existence of minds in the first place: according to such people, there are only brains, and the only acceptable way of describing the activity and condition of human beings is in the language of science; descriptions which use language appropriate to minds rather than brains (for example, talk about 'beliefs') should simply be abandoned as misleading and false. This approach to the mind-body problem will be our topic next week, as it raises substantial philosophical issues of its own; for now we'll proceed on the assumption that it *is* legitimate to discuss the human mind using our usual vocabulary of belief, desire, and sensation.

Defining our terms

When we ask 'What are minds?' this is often understood as a *metaphysical* question: are minds somehow made up of the same physical material (quarks, electrons etc.) which makes up the brain, or are minds made up of a radically different kind of stuff (traditionally referred to as '**mental substance'** to contrast it with the 'physical substance' that makes up our brains and bodies)? But there is also a *conceptual* question which we need to answer before we do any metaphysics: what do we *mean* by a 'mind', and in particular what distinguishes a **mental state** (for example, believing that Paris is the capital of France) from a **physical state** such as, say, being in Paris. Hopefully, doing this will enable us to distinguish **mental properties** from **physical properties**, and enable us to understand more precisely what we are talking about when we talk

about 'the mind': the mind is whatever it is that is the bearer of mental properties and the possessor of mental states. This approach to the mind is at least as old as Descartes (1596-1650):

> 'What am I, then? A thing that thinks. What is that? A thing that doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions.'
>
> Descartes, *Second Meditation*

Descartes defines himself (and hence his mind) in terms of a range of mental states and activities; thus to understand what we mean by a 'mind' we would need to say what all of these states have in common. The question of what distinguishes mental states (or properties) from physical ones is a controversial one. Here are some options:

1. First-person accessibility and incorrigibility.

A common thought is that we have some kind of special access to the contents of our own mind through a faculty of 'introspection', and that this gives us a special *authority* with regard to our own mental states: if I honestly believe I am in pain, no-one else has the right to contradict me. It is often thought that this special authority derives from the fact that our access to our own mental states is somehow *incorrigible* or *infallible* – it is logically impossible for us to be mistaken about the contents of our own minds.

One problem with this idea is that our special authority is limited only to mental states we are in *now*: I might sincerely believe that I was in great pain last Tuesday yet be mistaken (suppose I had forgotten that the accident occurred on Wednesday); and you might have every right to contradict me when I make such a claim, since your memory of events is not addled by trauma as mine is. So we have to limit the claim about our 'authority' to mental states we are in *now*. But is it even true that we have introspective access to our *current* mental states in such a way that we can define what it is to be a mental state in terms of availability to introspective access? In other words, can we say that a 'mark' of something's being a mental state is that, whenever one is suitably alert and conceptually sophisticated, one is in a position to know that one is in that mental state? This suggestion is considered, and rejected, by Tim Williamson:

> 'For example, one is sometimes in no position to know whether one is in the mental state of hoping [that] *p*. I believe that I do not hope for a particular result to a match; I am conscious of nothing but indifference; then my disappointment at one outcome reveals my hope for another. When I had that hope, I was in no position to know that I had it... [the suggestion] is even doubtful for the state of being in pain; with too much self-pity one may mistake an itch for a pain, with too little one may mistake a pain for an itch.' Williamson, *Knowledge and its Limits*, p.24

Indeed, it is easy to come up with examples of mental states which are not available to us until after the fact: suppose that at the time my teaching career seems like a succession of stressful and challenging tasks, and it is only later that I realize that I have enjoyed it tremendously. In 1902 Bertrand Russell experienced this dispiriting failure of introspection:

> 'I went out bicycling one afternoon, and suddenly, as I was riding along a country road, I realized that I no longer loved Alys. I had no idea until this moment that my love for her was even lessening.'
>
> *Russell*, Autobiography, p.147

If Russell reports the facts correctly, this is an example of a current mental state (his love for – or indifference

to – his first wife Alys) to which – for a while at least – he did not have introspective access, since up to the fateful bicycle ride he had been falling out of love with her without realizing it. In fact, the idea that sometimes we might not be authorities on our own state of mind is a staple of drama and literature:

(*Viewing*: Riley finds out he loves Buffy: BtVS Season 4 episode 7, 'the initiative')

2. Consciousness and qualia

(*Viewing*: *Star Trek the Next Generation* Season 2, Episode 9: *The Measure of a Man*)

It is a widespread thought that what makes us special is the fact that we have *consciousness* or *conscious experience*. But what is consciousness? Perhaps the most influential account is that given by Thomas Nagel in his 1974 article *What's it Like to be a Bat?*

> 'Conscious experience is a widespread phenomenon. It occurs at many levels of animal life, though we cannot be sure of its presence in the simpler organisms, and it is very difficult to say in general what provides evidence of it... But no matter how the form may vary, the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism... fundamentally an organism has conscious mental states if and only if there is something that it is like to *be* that organism – something it is like *for* the organism. We may call this the subjective character of experience.'

According to Nagel, beings are conscious if they have experiences with a 'subjective character' - experiences which feel a certain way. Subsequent philosophers have used the term **qualia** (singular **quale**) to refer to the 'subjective' or 'phenomenal' character these experiences have – the property of feeling a certain way. Here's how the *Stanford Encyclopedia* defines qualia:

> 'Consider your visual experience as you stare at a bright turquoise color patch in a paint store. There is something it is like for you subjectively to undergo that experience. What it is like to undergo the experience is very different from what it is like for you to experience a dull brown color patch. This difference is a difference in what is often called 'phenomenal character.' The phenomenal character of an experience is what it is like subjectively to undergo the experience. If you are told to focus your attention upon the phenomenal character of your experience, you will find that in doing so you are aware of certain qualities. These qualities — ones that are accessible to you introspectively and that together make up the phenomenal character of the experience are standardly called 'qualia'.

Although some philosophers (for example Daniel **Dennett**) deny that qualia exist, most accept that at least some mental states (for example hearing music, being tired, bored, or in pain) have qualia, as these are 'experiential' states – ones which involve the subject having an experience that *feels* a certain way. Galen **Strawson** has gone further, proposing the view that *every* mental state has qualia, since every genuinely mental state is 'experiential'. For example, he claims that there is such a thing as the *experience* of understanding a sentence or the *experience* of thinking of something.

We can probably agree that consciousness is a *sufficient* condition of being a mental state: if something is a conscious state (i.e. there is something it is like to be in that state), then it is a mental state. Pain is obviously a mental state, because being in pain *feels* a certain way. But if consciousness really is the distinguishing feature

of mental states then we need to be able to say that it is also *necessary* for being a mental state – i.e. something can be a mental state *only if* it is a conscious state; there are no such things as 'unconscious mental states'. That seems implausible: for one thing, since Freud it has been commonplace to recognize that our actions are sometimes caused by unconscious beliefs and desires which we may never even become aware of. For another, it is simply impossible for us to be conscious of all of our mental states at once: what would it be like consciously to consider *all* of your beliefs at the same time? At most, it seems that we are conscious of only a limited range of our mental states at once.

Nevertheless, John **Searle** has defended a version of this view, arguing that we understand what it is for something to be a mental state only because we understand what it is for something to be a conscious state, even in the case of unconscious mental states:

> '... there is no way to study the phenomena of the mind without implicitly or explicitly studying consciousness. The basic reason for this is that we really have no notion of the mental apart from our notion of consciousness. Of course, at any given point in a person's life, most of the mental phenomena in that person's existence are not present to consciousness ... however, we have no conception of an unconscious mental state except in terms derived from conscious mental states.'
>
> 'we understand the notion of an unconscious mental state only as a possible content of consciousness, only as the sort of thing that, though not conscious, and perhaps impossible to bring to consciousness for various reasons, nonetheless is the *sort of thing* that could be or could have been conscious.'
>
> Searle, *The Rediscovery of the Mind* (1992), pp.18-19; 155-6

3. Intentionality

Intentionality is not mentioned on the specification (so you can relax about this bit when revising), but it *is* important to know about if you want to understand why people find mental states so mysterious and hard to explain. Take a standard example of a mental state: my belief *that Paris is the capital of France*. That belief has a very distinctive feature: there is something the belief is *about*, namely Paris (and indeed France). This 'aboutness' or 'directedness' is called **intentionality** by philosophers. Many of the contents of my mind exhibit intentionality in the sense that they are **about or directed towards something in the world.** So your concept *dog*, for example, is about or directed towards actual dogs. And your beliefs in general are 'about' or directed towards things in the world because they are representations of how things are in the world. (Wittgenstein at one point tried to explain beliefs as being like 'pictures' of situations in reality.)

One thing that is strange about intentionality is that it is hard to see how *physical* objects could exhibit it. Suppose a neurosurgeon operates on your brain while you are thinking about the fact that Paris is the capital of France: could she find any part of your physical brain which was 'about' Paris? If she did, what physical properties would make it true that this part of your brain is about Paris rather than about Timbuktu? For that reason, many philosophers have argued that intentionality is a major problem for anyone who wants to explain minds in terms of brains.

For discussion: do *all* mental states exhibit intentionality? Can you think of any that don't?

More practice with terminology

By now you'll have noticed that there is quite a lot of terminology to get to grips with in this topic. Here's a quick summary:

Physical: this adjective is applied to anything which is part of the **physical world**, i.e. anything that has a physical location and interacts causally with other physical things. Your brain, for example, is a physical object. Some people have a different definition of 'physical': according to them, something is a physical object if it is part of one or another scientific theory put forward by physicists, biologists, or chemists. Quarks and electrons are physical objects because they are described by the science of physics.

NB: **Physicalism** can be defined as the doctrine that everything there is, is physical.

NB2: early modern philosophers used the phrase 'material object' where we would say 'physical object'.

Mental: the adjective 'mental' can be applied to anything that is part of a *mind*. So you can have mental abilities, capacities, states, properties, events and so on. Make sure you are confident distinguishing between these. A **mental state** is a *way your mind is* over a continuous period of time – for example, your **beliefs** are usually counted as mental states as they persist over a period of time. By way of contrast, things that happen in your mind at a particular time are usually counted as **mental events**; and **mental properties** are features or aspects of particular mental states or events. For example, it is a feature (property) of some of your beliefs that they are about Paris, and it is a feature (property) of the pain you feel after stubbing your toe that it feels exactly the way it does. An example of a mental ability or capacity would be your ability to remember the six times table.

Exercise:
Give your own examples of the following:

1) A mental state
2) A physical event
3) A mental event
4) A physical property
5) A mental ability

Homework essay: 'Is the mind the same as the brain?'
Remember to consider reasons on both sides of the debate, and to reach a clear evaluative conclusion. You haven't studied the material in this topic yet, so try to come up with some original thoughts of your own!

Week 2: Eliminative Materialism

'Eliminative' vs. 'Reductive' materialism

Materialists believe that there is only one kind of stuff in the world: physical matter, as described by scientists. Thus materialism can be defined as the view that **the mind is not ontologically distinct from the physical** – in other words, that your mind is not an *extra* thing which exists over and above your physical body. In contrast, **substance dualists** believe in two kinds of stuff: both physical matter and 'mental substance', which is the distinctive, non-physical kind of stuff that minds are made of. It is important to realize that **materialism** by itself is not a complete theory of the mind; there are many different materialist theories of mind to choose from.

An important distinction within materialism is between **reductive** and **eliminative** approaches: Materialists who take a **reductive** approach argue that all facts apparently about minds are really facts about the physical world 'in disguise'; in particular they might say they are facts about the neurobiological condition of the brain. They say that we could, in principle, 'derive' or work out all the facts about minds from facts stated in the language of science. Thus reductionists believe that the vocabulary we use in statements about minds (e.g. statements about a person's 'beliefs' or 'desires') is in principle dispensable: anything we might want to say in this language can in fact be stated using the language of science (even if we haven't yet worked out precisely how to do it). Crucially, although statements about minds are *dispensable* according to the reductionist, these statements are not *false.* In fact, the reductionist says, it is the very fact that statements about minds *can* in principle be translated into the language of science that shows how it is possible for statements about minds to express truths: they are true because of the truth of the scientific statements to which they are 'reducible'.

Eliminative materialists, on the other hand, take a harder line. They agree with the reductive materialist that the scientific facts are all the facts there are; but they deny that statements about minds can be 'reduced to' statements about brains. Thus they claim that our everyday talk about minds does not, in fact, express truths at all: concepts such as 'belief' and 'desire' should be abandoned as part of a

> 'radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience.'
>
> Churchland, *Eliminative Materialism and the Propositional Attitudes*

By 'ontology' here Churchland means the things which common-sense theories of the mind takes to *exist*: most notably mental states such as belief. We might have the *concept* of belief, but this fails to pick out anything real within our mind. Therefore eliminative materialism can be defined as the view that **some or all mental states do not exist**.

Eliminative materialists use the phrase '**folk psychology**' to talk about the system of concepts we use to describe others' (and our own) minds, and predict their behaviour. According to them, this system is as outdated (and false) as the theory that thunder is caused by the anger of Zeus; Although they concede that folk psychology enables us to make some rough-and-ready predications of other people's behaviour, it does not tell us about the *real* neurological causes of their actions, and we can confidently expect it to be replaced as soon as neuroscience has advanced to the point where it can give a complete explanation of human behaviour.

Here are some of the claims eliminative materialists make about folk psychology:

- Folk psychology is a *theory*, in the sense that it attempts to predict and explain observed phenomena by positing general laws which govern those phenomena. Like scientific theories, it has its own set of theoretical concepts: beliefs, desires, hopes, fears, intentions, decisions, perceptions, sensations, memories, etc. One example of a 'general law' which is part of folk psychology might be: if a person desires outcome x (all things considered), and believes that doing y will bring about x (without any negative side-effects), then that person will do y. This 'law' attempts to predict action using concepts of 'belief' and 'desire'. (The claim that we use a theory to make sense of other people's behaviour is called the **theory-theory**, since it is the theory that our understanding of other people is itself based on an empirical theory.)
- Since folk psychology is a theory, our acceptance of it is *conditional* on it doing the best job of explaining the observed phenomena; if we find that it conflicts with the observed phenomena, or that we have another theory that does a better job of explaining these phenomena, we are rationally obliged to abandon folk psychology, and also to abandon the system of concepts it contains.
- In the history of science, replacing one theory with another has frequently required us to stop believing in entities or properties presupposed by the old theory. For example, we used to explain mental illness in terms of demonic possession, heat in terms of 'caloric fluid', combustion and oxidation in terms of 'phlogiston'. Because those theories have now been abandoned, so too has belief in demons, phlogiston, and caloric fluid. Likewise, once we give up the theory of folk psychology, we should no longer believe that there are such things as 'beliefs' and 'desires'.
- The thesis of the **causal closure of the physical**. Since every physical event has a sufficient *physical* cause, it follows that the real causes of my actions must be physical events such as neuronal activity in my brain. Therefore, mental events such as decisions cannot be among the causes of my actions and so a theory which describes mental events, as folk psychology does, cannot give the *real* explanation of my actions.

So the eliminative materialist has a coherent account of the status of our talk about minds; but why should we believe their claim that this 'folk psychological' theory is radically false and so should be abandoned? Here are some of the main **arguments for eliminative materialism**:

- The limitations and shortcomings of folk psychology in providing explanations of human behaviour:

  > 'As examples of central and important mental phenomena that remain largely or wholly mysterious within the framework of FP, consider the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individuals. Consider our utter ignorance of the nature and function of sleep ... Reflect on the common ability to catch an outfield fly ball on the run, or hit a moving car with a snowball... Or consider the miracle of memory, with its lightning capacity for relevant retrieval. On these and many other mental phenomena, FP sheds negligable light.' (Churchland, *Op. Cit.*)

- The historical failure of other theories based on 'common sense' - for example, 'folk physics', according to which every object in the universe tends to fall in the same direction (down), and heavier

objects fall faster. If common sense can get it wrong so radically when it comes to explaining the motion of physical objects, why should we expect it to be reliable when it comes to explaining the behaviour of human beings?

- If we recognize that folk psychology is a (falsifiable) theory aimed at explaining human actions, we can provide answers to several problems in the philosophy of mind. For example, there is no longer a 'mind-body' problem, since the theory that implies there are minds (i.e. folk-psychology) can be rejected as false and misleading. Moreover, it offers a neat solution to the 'problem of other minds':

  > 'The problematic conviction that another individual is the subject of certain mental states is not inferred deductively from his behaviour, nor is it inferred by inductive analogy from the perilously isolated instance of one's own case. Rather, the conviction is a singular *explanatory hypothesis* of a perfectly straightforward kind. Its function, in conjunction with the background laws of folk psychology, is to provide explanations / predictions / understanding of the individual's continuing behaviour, and it is credible to the degree that it is successful in this regard over competing hypotheses.' Churchland, *Op. Cit.*

Reading: Churchland, *Eliminative Materialism and the propositional attitudes*, s.I&II (pp.67-76). Label paragraphs in green, orange, and red, depending on how well you feel you have understood them.

Study questions:

1) What does Churchland mean by saying that in primitive cultures, 'the behaviour of most of the elements of nature were understood in intentional terms'?
2) What do you understand by the claim that folk psychology is a 'degenerating research programme'? Do you agree? Why?

Arguments *against* eliminative materialism.

The emergence of a philosophical theory according to which there are no such things as beliefs, desires, intentions, and feelings provoked exactly the level of hostile reaction you would expect. Here we should focus on four of the most important arguments against eliminative materialism: that we can *know* it to be false on the basis of our direct awareness of mental states or qualia; that it is self-refuting; that our understanding of others in terms of intentional states such as belief and desire is *not* based on a theory of 'folk psychology'; that the success of folk psychology at predicting the behaviour of others gives us reason to believe that at least *some* folk-psychological concepts will have lasting relevance.

Direct awareness

One serious problem for the eliminative materialist is that we seem to be directly aware of examples of the mental states they claim do not exist. For example, introspection allows me to know that I *believe* that I am in this room, and to know that I *want* more coffee. So my view that beliefs and desires exist is not 'theoretical', based on my attempt to understand other people's behaviour; instead I have direct *access* to my own mental states, and so can know that such intentional mental states exist. At any rate, I have more reason to believe that these mental states exist than I have to trust any argument that claims they do not. This could be put by saying that **the intuitive certainty of the existence of my mind takes priority over other considerations**: I cannot coherently doubt whether I am thinking (remember Descartes' *Cogito* argument from last year? It follows from the fact that I am doubting whether I am thinking that I actually *am* thinking, since doubt is a kind of

thought.). So the claim that I am thinking – and therefore have mental states – is certain; moreover it is certain simply as the result of rational 'intuition': it is a deliverance of what Descartes called the 'natural light' of reason.

In response to this, the eliminative materialist can say that attributing belief- and desire-states to ourselves can only be done *within* the theoretical framework of folk psychology, and that my self-attribution of beliefs and desires comes about because I interpret my introspective evidence in line with the folk-psychological theory I already have; if I were not in the grip of the folk-psychological way of thinking about myself, I would not interpret even myself as having beliefs.

However, it is not clear that this response is entirely successful. What about our awareness of mental states which subjectively *feel* a certain way – i.e. states with qualia, such as being in pain? According to the eliminative materialist, our conscious awareness of pain is not part of the real explanation of why we respond the way we do to certain stimuli, since the real explanation takes place at the neurological level. Consciously felt pain may be just another element of the soon-to-be discredited folk psychological theory of human action, and does not really exist. But surely our awareness that we are conscious of pain is *not* 'theoretical' - rather, pain is a mental state to which we have *infallible* access, and so we have at least one example of a mental state whose existence is not in principle falsifiable, and so not part of a 'theory' which we could one day reject.

Nevertheless, the eliminative materialist has at least three responses to this argument. First, she can accuse us of committing the '**phenomenological fallacy**': a term coined by Ullin **Place** in the 1950s for the alleged mistake of supposing that whenever it appears to us that things are a certain way, there is some entity within our mind that really is that way – for example, supposing that when we experience pain, this is because there really is something – a painful mental state – which is the thing which we experience. As Place puts it, it is

> 'the mistake of supposing that when the subject describes his experience, when he describes how things look, sound, smell, taste, or feel to him, he is describing the literal properties of objects and events on a peculiar sort of internal cinema or television screen, usually referred to in the modern psychological literature as the 'phenomenal field'.'
>
> Place, *Is Consciousness a Brain Process?*

If Place is right, it is a mistake to argue from the agreed fact that people sometimes are in pain, to the conclusion that there is a such a thing as the pain they experience; the experience of pain does not entitle us to infer the existence of painful mental states.

A second response to the problem pain poses to the eliminative materialist is to point out that our idea of pain is confused and even potentially contradictory – making it sound like a concept that is part of our 'common-sense' folk psychological system, and in need of replacement by a more coherent concept from advanced neuroscience. Here the eliminative materialist will point to the phenomenon of **reactive disassociation** – a situation apparently found in patients who have undergone lobotomies or received morphine treatment for acute pain. Such people report that the pain is still present, and they are still aware of it – but that it does not bother them any more; although they are genuinely in pain, this does not strike them as something unpleasant. Such examples present a dilemma for the common-sense conception of pain: either we are mistaken in believing that pain is (by definition) something that feels intrinsically awful, or *they* are mistaken in claiming

to be in pain when they are not – but then our own claim to be infallibly aware of whether we are in pain is undermined, and with it the claim that we can be sure that pain is one mental state which cannot be eliminated.

Finally, the eliminative materialist might adopt a position according to which qualia such as those associated with pain are not eliminated, but rather reduced: shown to be identical with certain brain functions. This would be a *partially* eliminative materialism, as it would claim that, while most mental states (e.g. those involving belief and desire) do not exist, some 'qualitative' mental states such as pain, do exist, but can be *explained* in materialist, neuroscientific terms.

The Self-Refutation Objection

Several philosophers have argued that **the articulation of eliminative materialism as a theory is self-refuting**. A short way of putting the criticism is this: if it is true that beliefs do not exist, then no one can believe the claims of eliminative materialism. By making claims about what we should and shouldn't believe, the eliminative materialist is presupposing the existence of the very entities (beliefs) whose existence she denies. Put in this way, the argument invites a simple response: the eliminative materialist can say that her claim 'beliefs and desires do not exist' is not intended to alter people's (non-existent) beliefs, but rather to change their behaviour (to make them stop uttering the sentences that express commitment to the existence of beliefs); and no-one is trying to argue that human *behaviour* does not exist.

A more sophisticated version of the argument may evade this criticism, however. To be able to state the hypothesis of eliminative materialism in the first place, the theorist must be able to use sentences *meaningfully*, to assert propositions with a definite *content*. But it seems plausible that assertion is possible only when there is *intention* to communicate, and *knowledge* of what content is being expressed, and according to the eliminative materialist 'knowledge' and 'intention' do not really exist, as they are artefacts of the discredited theory of folk psychology. Lynne Rudder Baker has defended this argument:

> 'language can be meaningful only if it is possible that someone mean something. This is a platitude, not a theory. It is clearly incumbent upon anyone who wants to deny the platitude to show that there can be meaningful language even if no-one has meant anything, even if no-one has ever intended anything.' Baker, *Cognitive Suicide*

Rejecting the Theory-Theory

A cornerstone of the eliminative materialist proposal is the so-called **theory-theory**: the theory that our application of terms such as 'belief' and 'desire' is part of an overall theory of human behaviour called 'folk psychology', which is in principle falsifiable if new evidence comes to light, and which should be rejected once we have a new theory which does a better job of predicting human behaviour. But it is possible to argue that folk psychology is not an *optional* way of thinking about other people, which we could (and should) reject, but rather is something we are obliged or constrained to do by our nature as human beings, and which we could not abandon so long as we continue to treat other people *as* people.

This approach was developed by P.F. Strawson: he claimed that ceasing to recognize someone as motivated by beliefs and desires and thus susceptible to rational argument, would require us to adopt the '**objective attitude**' towards them, and this is something that is neither desirable nor possible:

> 'To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided ... The objective attitude ... cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other... A sustained objectivity of inter-personal attitude, and the human isolation which that would entail, does not seem to be something of which human beings would be capable, even if some general truth were a theoretical ground for it. '
>
> Strawson, *Freedom and Resentment*

If Strawson is right, thinking about other people as motivated by beliefs and desires, and swayed by reasons, is a necessary condition for treating them *as* people in the first place; and abandoning this stance towards them is not a genuine possibility, however much neuroscience might develop.

The 'Success' of Folk Psychology

We should notice a strange fact about eliminative materialists: they may be willing to accept that, at the moment, folk psychology is the *best* theory we currently have for predicting and explaining human behaviour; however, they claim that reflection on the nature and status of that theory enables us to be reasonably sure that the theory will one day be replaced by something genuinely 'scientific' rather than based on 'common sense'. But this means that their theory is **hostage to fortune**: whether or not they are right will depend on how things turn out in the sciences; in particular, on whether a completed neuroscience does a better job of explaining human behaviour than folk psychology does.

Moreover, it may be that, when we finally do develop a neuroscientific theory which does a better job of explaining human action than folk psychology does, this does not result in us abandoning all of the concepts within our original folk psychological theory, or concluding that these concepts do not represent anything real. Although the history of science provides examples of instances where changes in theory require old concepts to be abandoned all together (as in the case of phlogiston), there are also examples of concepts being preserved even though the theory surrounding them has changed radically. For example, the switch from geocentric to heliocentric theories of the solar system did not result in us abandoning the concept of a planet. Defenders of folk psychology can use examples like this to argue that the most likely outcome is that everyday mental concepts such as belief and desire will be revised and reworked in the light of scientific discovery, without being abandoned altogether. This position is known as **revisionary materialism**: the view that scientific advances will prompt a partial revision of our common-sense theory of mind, rather than its total abandonment as the eliminative materialist predicts.

In defence of such a view, the opponent of eliminative materialism will appeal to the apparent success of folk psychology in predicting human behaviour: most of the time, we can make accurate predictions about what people will do next on the basis of attributing beliefs and desires to them, and this at least provides us with a **pragmatic justification** for using folk psychology: a justification in terms of 'what works'. You might argue thus: **folk psychology has good predictive and explanatory power**; therefore, unless someone comes up with

a better theory (one which explains more, and predicts future events with more accuracy), it would be unjustified and irrational not to believe that folk psychology is actually *true*.

Nevertheless, the eliminative materialist will respond that optimism about the survival of folk psychology on the basis of its predictive and explanatory power is simply the result of ignorance of the history of science: the mere fact that a theory appears successful in predicting the things we use it to predict does not show that it accurately represents the way things really are. We can use Newtonian mechanics to predict the movement of physical objects quite accurately even while treating those objects as continuous solids whose mass is evenly distributed throughout; but of course the model of reality we rely on in making such calculations is radically misleading, not only because it fails to recognize that 'objects' are in fact collections of molecules rather than continuous solids, but also because it incorporates fundamental errors about the nature of gravity and space-time itself. The eliminative materialist will say that folk-psychology is in exactly the same situation as Newtonian mechanics: it makes broadly accurate predictions, but that gives us no reason to suppose it accurately represents the way things really are.

Week 3: Substance dualism

It is a striking fact about minds that their properties are not ones we attribute to anything in the physical world: consciousness, intentionality, self-awareness, rationality and freedom are not qualities that we are ordinarily tempted to apply to physical objects. This provides a *prima facie* reason for thinking that minds themselves are not physical entities (or 'substances'); that instead there must be *two* kinds of substance in the world: physical stuff, ('material substance'), and mental substance. Physical objects are material substances, while minds are mental substances, and consequently do not occupy any space in the physical world, and are not subject to the laws of physics. This view is known as **substance dualism**, or sometimes **Cartesian dualism**, in honour of the fact that its most high-profile defender was **Descartes**. In fact, when discussing arguments for substance dualism, you'll most often be talking about Descartes' arguments for this view, so it is worth going into them in detail.

A note about definitions: **dualism** can be defined as the view that **the mind is distinct from the physical**. There are actually two ways of being a dualist. The most straightforward is Descartes' **substance dualism**, according to which the mind is a separate, non-physical substance: the mind is a different kind of thing from any physical object. A more complex view is **property dualism**. This claims that, although there is not an independently existing thing which we might call the 'mind', each of us has **mental properties** which cannot be explained in terms of, or deduced or figured out from, physical properties of our brain. According to the property dualist, although there are not two kinds of *thing* (substance) in the world, there are two radically different kinds of *property*. One very common way of making this claim is to say that **qualia** (the properties of our experiences that make them feel they way they do) are a distinctive kind of non-physical property that cannot be explained in terms of anything physical. For example, the way pain feels to you is a non-physical property of your experience.

Earlier we said that the alternative to dualism is **materialism**. This is sometimes defined (for example, in the A-level specification) as the view that **the mind is not ontologically distinct from the physical**. But that definition now needs some fine-tuning, since materialism is not just opposed to substance dualism (which claims that the mind is a separate, ontologically distinct, non-physical entity); materialists *also* reject property dualism. So materialism is not merely a view about the mind (that the mind is not a separate non-physical substance), but also about mental properties and mental states: that these are not separate, 'ontologically distinct' kinds of thing. Better, then, to define materialism as the view that minds and mental properties are not ontologically distinct from physical objects and properties; there is nothing that is not physical. (Remember: *eliminative* materialists say that minds and mental states don't exist in the first place!)

In more recent debate, philosophers often describe themselves as 'physicalists' rather than 'materialists' to emphasize their commitment to the view that there is *nothing* in the world which is not physical: the world contains nothing except what a complete physical science would say that it contains (by 'physical science' philosophers mean not only physics, but also chemistry, molecular biology, and neuroscience). For our purposes we can treat physicalism and materialism as different names of what is fundamentally the same view.

Exercise: Make a diagram showing the relationship between materialism and the different kinds of dualism.

Reading: Descartes, *Sixth Meditation*

Two of Descartes' arguments for substance dualism

Earlier (in the *Second Meditation*), Descartes had considered the suggesting that the mind and body must be distinct because he could doubt the existence of his body, but could not doubt the existence of his mind. This is because he can imagine a situation in which he is being supplied with images of a physical world (by an Evil Demon) which make him believe that he has a physical body, but where in fact there is no physical world whatsoever, and he exists just as a mind being continuously deceived. Thus Descartes can claim that, although he knows that his mind exists, he does not know that his body exists – so the two must be different. This is called Descartes' **knowledge argument** by philosophers, but is not included in the specification (possibly because it is such a bad argument that there are signs that even Descartes himself did not want to rely on it). The flaw is this: you cannot reason 'I know *x* but I do not know *y*, therefore *x* is not *y*.' Try thinking it through substituting Clark Kent and Superman for *x* and *y* and you'll soon see why it's a logical fallacy.

The conceivability argument

Descartes' main reasons for supposing mind and body to be distinct substances are presented at the end of his work, in the *sixth* meditation. The first of these, the **conceivability argument**, relies on the **logical possibility of mental substance existing without the physical**. In other words, Descartes first argues (1) that it is conceivable for his mind to exist without being 'embodied' in a physical world, then claims that, as a result, (2) it is logically (or metaphysically) possible that his mind could exist without a physical body (i.e. that it *really could have been the case* that he existed without a body), then draws the further conclusion (3) that his mind and body are 'separate and distinct' substances. The move from (2) to (3) is a result of Descartes' definition of a 'substance': a substance is something that does not depend on anything else to exist. For the mind and body to be 'separate substances', it does not have to be the case that the mind actually *does* exist without a body – all that is needed is that the mind *could have* existed without a body. For example: the body and the lid of my pen are separate substances, because they *could* exist even if they were separate, and one of them could survive the destruction of the other. It wouldn't matter if, in fact, no-one ever bothered to separate them in this way; they would still be 'separate substances' merely as a consequence of the fact that they *could* have existed without each other.

Exam tip: a frequent mistake in the old Descartes set text paper was for students to say that Descartes' argument fails 'because he only proves that the mind and body *could* be separated, not that they actually *have* been'. Hopefully you can see why this is a howler: it does not appreciate the fact that, in Descartes' system, merely being *able* to be separated (not depending on each other to exist) is enough for the two to be genuinely 'separate substances'. Try to avoid making this mistake in the exam.

So the big questions for the conceivability argument are, first, why does Descartes think that he can coherently conceive of his mind as existing without a body, and second, why does Descartes think that **what is conceivable is possible?** Read the extract below and see if you can work out how Descartes would answer those two questions:

> 'And, firstly, because I know that all which I clearly and distinctly conceive can be produced by God exactly as I conceive it, it is sufficient that I am able clearly and distinctly to conceive one thing apart from another, in order to be certain that the one is different from the other, seeing they may at least be made to exist separately, by the omnipotence of God; and it matters not by what power this separation

is made, in order to be compelled to judge them different; and, therefore, merely because I know with certitude that I exist, and because, in the meantime, I do not observe that aught necessarily belongs to my nature or essence beyond my being a thinking thing, I rightly conclude that my essence consists only in my being a thinking thing or a substance whose whole essence or nature is merely thinking]. And although I may, or rather, as I will shortly say, although I certainly do possess a body with which I am very closely conjoined; nevertheless, because, on the one hand, I have a clear and distinct idea of myself, in as far as I am only a thinking and unextended thing, and as, on the other hand, I possess a distinct idea of body, in as far as it is only an extended and unthinking thing, it is certain that I, that is, my mind, by which I am what I am], is entirely and truly distinct from my body, and may exist without it.' Descartes, *Sixth Meditation*

Commentary

Descartes' presentation includes these claims:

- I have separate 'clear and distinct' ideas of mind and body: I conceive of myself as a 'thinking and unextended thing', and of my body as an 'extended and unthinking thing'.
- God has the power to produce things which I clearly and distinctly conceive 'precisely as I conceive them'. This follows from God's omnipotence: if God can do anything that is not contradictory, then God can bring about situations which I clearly and distinctly conceive, since these situations are guaranteed to be not contradictory by the mere fact that I can 'conceive' them (represent them to myself) without confusion or incoherence.
- Therefore, God has the power to make my mind exist separately from my body, and so my mind *can* exist (it is 'logically possible' for my mind to exist) without a body;
- Therefore, my mind is 'separate and distinct' from my body.

To answer the questions posed earlier: Descartes' reason for saying that he can coherently conceive of the mind as existing without the body is that he can 'clearly and distinctly' conceive of the mind as just a 'thinking thing' lacking a physical body, and this is all the reason he needs for believing that it is genuinely conceivable that the mind should exist in that way. Secondly, Descartes bridges the gap from *conceivability* to *possibility* by appealing to God's omnipotence: it might be *conceivable* that I could exist without a body, but that does not automatically prove that it is *possible* (it might be a logical impossibility for there to be non-physical substance, or I might in fact be identical with my brain). God's omnipotence guarantees that the mind *can* be separated just because Descartes can conceive of it being separate, since God (as an omnipotent being) can do everything conceivable. And of course, if the mind *can* be separated in principle, then the mind does not depend on anything else for its existence, and so counts as a 'separate and distinct substance' by Descartes' own definitions.

Criticism of the conceivability argument

First, notice that the argument (as Descartes states it) relies on the existence of an omnipotent God; but if (as many think) Descartes' attempts to prove the existence of God are unsuccessful, then he does not know that God exists and cannot use the existence of God as a premise in his argument. Some critics express doubt about whether it is possible for the conceivability argument can be **expressed without reference to God**. In other words, we might suspect that, without God supplying the necessary justification for the move from conceivability to possibility we should not infer the latter from the former. And in fact, we might have independent reasons for thinking that **what is conceivable is not always possible**. One example of this was provided in the *Objections* to the *Meditations* by the French philosopher Antoine **Arnauld**: he suggests that someone could form a conception of a right-angled triangle while doubting that Pythagoras' law is true, and that such a person could mistakenly argue that, because he could *conceive* of a right-angled triangle where the square of the hypoteneuse is not equal to the sum of the squares of the other two sides, it must therefore be possible for such a triangle to exist – although of course, it is *not* possible for such a triangle to exist. Descartes spent some time responding to this issue, but in short his main point was that the man in the example did not have a 'complete' conception of the triangle, and it is only complete conceptions of the nature of a thing that enable us to draw conclusions about what is possible. (You might respond: how is Descartes so sure that his

conception of the mind as a 'thinking thing' is after all a 'complete' conception?)

The move from conceivability to possibility (without appealing to God's existence) is justified by David Chalmers with reference to his 'Zombie' argument, which we'll encounter later in the course:

> 'Arguing for a logical possibility is not entirely straightforward. How, for example, would one argue that a mile-high unicycle is logically possible? It just seems obvious. Although no such thing exists in the real world, the description certainly appears to be coherent... I can discern no contradiction in the description... If no reasonable analysis of the terms in question points towards a contradiction, or even makes the existence of a contradiction possible, then there is a natural assumption in favour of logical possibility.' Chalmers, *The Conscious Mind*

Chalmer's point here is that we should assume that a situation that seems to have a coherent description *is* logically possible *unless* there is a good reason for thinking that there is a contradiction which can be revealed by analysis; but if there is no reason for thinking there is a contradiction then we should be happy to accept the claim to logical possibility. This is a kind of 'burden of proof' argument: Chalmers claims that there is a 'natural assumption' in favour of logical possibility, so the burden of proof lies on someone who thinks that a situation is logically *impossible* – they are the ones who have to provide an argument to support their view, not him.

A second objection to the conceivability argument is that Descartes' starting point is wrong: he thinks he can conceive of his mind existing without a body, but this is a mistake: in fact, a **mind without a body is not conceivable**. One way of arguing for this is as follows: to be a subject is to have sensory perceptions; in fact it is inconceivable that someone could exist without having sensory perceptions (remember, even in the Evil Demon case, the subject *has* perceptions: it's just that these are systematically misleading. But you can have perceptions only if you can make sense of them in terms of your own physical location (using the concepts of up, down, left, right, near, far and so on). So to conceive of yourself (or anyone else) as existing just is to conceive of yourself as occupying some physical location which enables you to make sense of your perceptions. Here's how Charles Taylor makes the point (he attributes the argument to the early 20th century philosopher Merleau-Ponty):

> 'To be a subject is to be aware of a world. I can be aware of the world in many ways. I can be pondering the situation in Namibia or last year at Marienbad, considering the second law of thermodynamics, and so on. But the one way of having a world which is basic to all this is my perceiving it from where I am, with my senses, as we say. This is basic, first because it is always there, as long as I am aware at all; and second because it is the foundation of other ways of having a world. We can ponder distant events, or theoretical perspectives on things, because we are first of all open to a world which can be explored, learnt about, theorized about, and so on. And our primary opening to this world, the inescapable background to all others, is through perception. Now our perception of the world is essentially that of an embodied agent, engaged with, or at grips with the world. And once again, the term 'essentially' carries the force discussed above; the claim is not just that perception depends causally on certain states of our bodies - that I couldn't see if my eyes were not in good condition, or the like. The claim is rather that our perception as an experience is such that it could only be that of an embodied agent engaged with the world. Let's consider. Our perceptual field has an

> orientational structure, a foreground and a background, an up and down. And it must have; that is, it can't lose this structure without ceasing to be a perceptual field in the full sense, i.e., our opening onto a world.'
>
> Taylor, *The Validity of Transcendental Arguments* (1978).

If this line of argument is successful, Descartes' own starting point is mistaken: it is not the case that he has a complete 'clear and distinct' conception of his mind as *just* a thinking thing; rather, he should conceive of himself as a thinking *and* perceiving thing, and to do that he must conceive of himself as a *located* thing, and therefore as something having a body.

A *third* objection to the conceivability argument is the suggestion that **what is logically possible tells us nothing about reality**. One way of developing this should be familiar from last year's work on **Hume**: you might think that judgements about logical possibility tell us only about the limitations of the ways we can think about the world, not about the world itself. This is because what we can and cannot conceive of tells us only about 'relations of ideas': these are conceptual truths about *our* way of thinking, but there is no guarantee that the world itself will conform to or 'fit with' our own human way of thinking about it. Another way of putting the point is to note that – even if it is logically possible for the mind to be a thinking and unextended thing – it is equally logically possible for the mind to be a thinking and *extended* thing. We want to know which of these is the conception of the mind that accurately represents the way things are in reality; but the mere fact of logical possibility will not enable us to choose between them.

However, Descartes can reply to this second version of the criticism. Showing that it is conceivable that the mind should exist without a body tells us something about reality by telling us something about the 'essence' or 'essential nature' of the mind: those features of the mind which are needed for something to be a mind in the first place. Specifically, from the fact that the mind is not essentially embodied, and therefore *can* exist without a body, it follows that the mind does not depend on the body for its existence; and of course by Descartes' definition of 'separate substance', if the mind does not depend on the body then it is a separate substance from the body after all.

Extension: the A2 specification only talks about *one* kind of possibility: logical possibility. Something is logically possible if and only if it does not violate the laws of logic. (So for example, it is logically *impossible* for you both to be eighteen years old and not to be eighteen years old, since if that were true it would violate the logical law of non-contradiction, that there are no true sentences of the form '*p* and not-*p*'.) But you might think that, for his argument to work, Descartes needs not only logical possibility (consistency with the laws of logic), but also metaphysical possibility. This is the kind of possibility we have where something really *could have been the case*. To show that his mind does not depend on his body for its existence, Descartes needs to show not only that the idea of his mind existing without a body is consistent with the laws of logic (logical possibility) but also that it really could have been the case that his mind existed while lacking a body (metaphysical possibility) – but this latter kind of possibility will be much harder to argue for. So by distinguishing between different kinds of 'possibility' we can arrive at a third, more substantial way of understanding the criticism that what is logically possible tells us nothing about reality: namely, that logical possibility is not the right *kind* of possibility to establish the conclusion Descartes intends concerning the separateness of the mind and body. Logical possibility tells us only about how the laws of logic work; what we

need to draw conclusions about the way the world could be is metaphysical possibility. (You might want to think about whether the appeal to God's omnipotence enables Descartes to answer this criticism.)

The indivisibility argument

Descartes' other argument for the separateness of mind and body relies on **Leibniz' Law** (sometimes called the principle of the **indiscernibility of identicals**): that one and the same object cannot have contradictory properties: if *a* is *F* and *b* is not *F*, then *a* and *b* cannot be the same object. To generalize, if we can say that the mind has properties which the body does not (or *vice versa*), then we can use Leibniz' law to conclude that the mind is not the same as the body. Here is Descartes' presentation of the argument:

> 'To commence this examination accordingly, I here remark, in the first place, that there is a vast difference between mind and body, in respect that body, from its nature, is always divisible, and that mind is entirely indivisible. For in truth, when I consider the mind, that is, when I consider myself in so far only as I am a thinking thing, I can distinguish in myself no parts, but I very clearly discern that I am somewhat absolutely one and entire; and although the whole mind seems to be united to the whole body, yet, when a foot, an arm, or any other part is cut off, I am conscious that nothing has been taken from my mind; nor can the faculties of willing, perceiving, conceiving, etc., properly be called its parts, for it is the same mind that is exercised all entire in willing, in perceiving, and in conceiving, etc. But quite the opposite holds in corporeal or extended things; for I cannot imagine any one of them how small soever it may be, which I cannot easily sunder in thought, and which, therefore, I do not know to be divisible. This would be sufficient to teach me that the mind or soul of man is entirely different from the body, if I had not already been apprised of it on other grounds.'

In other words:

1) The body is divisible into parts
2) The mind is not divisible into parts
3) Therefore, the mind and the body are different substances.

Criticism of this argument focusses on two areas: first, we might claim that Descartes' argument fails either because **not everything thought of as physical is divisible** or because **the mental is divisible in some sense**. To take the first of these: Descartes does indeed say that 'corporeal or extended things', no matter how small, can always be divided ('sundered') in thought even if too small to see. However small a physical object is, we can always imagine cutting it in half with each of the resulting halves taking up exactly half as much space as the whole. If Descartes is right about this, it follows that the mind cannot be the same as *any* physical object, as all physical objects are divisible, but the mind is indivisible. As he says in his 'Preface to the Reader',

> 'we cannot understand a body except as being divisible, while by contrast we cannot understand a mind except as being indivisible. For we cannot conceive of half a mind, while we can always conceive of half a body, however small; and this leads us to recognize that the natures of mind and body are not only different, but in some way opposite.'

However, we might object that modern physics provides examples of particles which – although part of the physical world – do not take up an *area* of physical space and so cannot be divided into parts in the way Descartes supposed. For example, electrons are often conceived of as 'point particles' which occupy exactly one point in space without having a physical size in any direction, and which are therefore indivisible: a

particle is divisible only if it has spatial parts, but it can have spatial parts only if it occupies an area of space, and electrons do not. Physicists will reply that this objection misunderstands the nature of the physics involved: the notion of point particles is an *idealization* for the purposes of certain calculations rather than the way things are in reality (for example, even a planet can be treated as a point particle located at its centre of gravity for some astronomical calculations). Worse still, in quantum mechanics it may be claimed that even elementary particles (the ones which are not made up of other smaller particles) have to be treated as taking up an area of space. However, this does not answer the philosophical issue. The problem for Descartes is that physics shows that we can make sense of the idea of a physical object – a point particle – that takes up no space whatsoever. Even if it turns out that all the actual physical particles do take up space, it doesn't matter: what we have shown is that it is *conceptually coherent* to imagine something that is both physical and indivisible, so Descartes is wrong to claim that it is part of the idea of something being physical that it takes up space and is divisible, so he cannot conclude, from the fact that the mind is indivisible, that the mind is not physical – in short, *being physical* does not entail *being divisible* as Descartes supposed.

A better response is possible. Descartes himself thinks that his indivisibility argument shows that the mind cannot be identical with *any* 'body' or physical object. If there can be physical objects which are indivisible 'point particles' then his claim is false. However, Descartes' argument does show that his mind is not identical with *his body* or indeed his *brain*, since both of these physical objects are divisible, and his mind is not. So the indivisibility argument can be used to show, not that the mind is not physical, but rather that the mind is not the same as the brain, which, after all is what most 'identity theorists' want to say (more on that in coming weeks). Moreover, it shows that the mind is not the same as any *part* of the brain, as long as that part is itself divisible. The only way the mind could be identical with anything within the brain would be for it to be identical with an indivisible point particle in the brain, such as one particular electron – but that is an absurd suggestion that no-one seriously wants to defend. So Descartes' argument still presents a challenge to theorists who want to say that the mind is the same as the brain.

The second objection to the indivisibility argument is that **the mental is divisible in some sense.** It might seem obvious that minds can in fact be divided into different parts – for example, memory and desire. Moreover, Freudian analysts treat the mind as consisting of ego, superego, and id. In Freud's theory, the ego attempts to balance out the competing drives supplied by the childlike id and overly moralistic superego. (Two thousand years earlier, Plato had proposed a very similar division of the mind into three competing parts.) If the mind, after all, *is* divisible, then Descartes cannot use Leibniz' law to show that it is distinct from the brain.

However, the view that the mind is *indivisible* has been defended by E.J.Lowe:

> 'the self patently does not consist of a plurality of lesser 'selves', acting co-operatively... Nor should we regard the mind's various 'faculties' - will, intellect, appetite, or modern variants thereof, such as linguistic or visual processing 'modules' - as being 'parts' of the self. For in the first place it is a mistake to reify faculties, and in any case they certainly could not qualify as *substantial* parts, which are what are now at issue. Faculties have no possibility of independent existence and should properly be seen as no more than abstractions from the mental lives of persons. For instance, the notion of a will without an intellect, or of a language faculty in the absence of belief and desire, is just nonsense.'
>
> E.J. Lowe, *Subjects of Experience*, ch.2

Lowe's view is that the 'self' cannot be divided into independent, 'substantial' parts; even if we choose to distinguish different functions or capacities within the mind, these are not things which could exist by themselves, and so the mind does not count as 'divisible' in the sense that an extended physical object is divisible. What makes a physical object count as 'divisible' is that it has parts which can survive being separated from one another, but the alleged 'parts' of the mind cannot survive separation in this way – they have 'no possibility of independent existence'. If that is right, then Descartes' argument can survive this second criticism.

Week 4: The Interaction Problem

The main problem for substance dualism is the **interaction problem** – the question of how (and if) the immaterial mind interacts causally with the physical brain. It certainly *seems* very much as though events in our minds cause changes in the physical world, and events in the physical world cause events in our minds. For example, my decision to go out (mental event) causes me to get up and put my coat on (physical events), while my drinking a cup of coffee (physical event) causes an increase in my overall alertness (mental event). Dualists who accept (as Descartes did) that there *is* this kind of two-way causal interaction between mind and body are called **interactionists**; the problem for such people is to give an account of how this kind of interaction is possible.

Descartes attempted to explain the interaction between mind and body by locating the part of the brain where it happened: the pineal gland, which, he claimed, contained 'animal spirits' which transfer the energy of the mind out into the rest of the brain.

> 'My view is that this gland is the principal seat of the soul, and the place in which all our thoughts are formed. The reason I believe this is that I cannot find any part of the brain, except this, which is not double. Since we see only one thing with two eyes, and hear only one voice with two ears, and in short have never more than one thought at a time, it must necessarily be the case that the impressions which enter by the two eyes or by the two ears, and so on, unite with each other in some part of the body before being considered by the soul. Now it is impossible to find any such place in the whole head except this gland; moreover it is situated in the most suitable possible place for this purpose, in the middle of all the concavities; and it is supported and surrounded by the little branches of the carotid arteries which bring the spirits into the brain' Descartes, *Letters*

This account has not met with general acceptance. For one thing, modern biology has established that the pineal gland functions to produce hormones such as melatonin, which controls the sleep-wake cycle, rather than transmitting the decisions of the mind to the brain. For another, mind-body interaction is not explained in any way merely by saying *where* it happens. In fact, we now know quite a lot about which areas of the brain are associated with which types of mental activity – but this does not amount to an increased understanding of *how* an immaterial mind might bring about events in the physical brain.

We can divide the **interaction problem** into **conceptual** and **empirical** issues to do with **causation**. First, take the conceptual difficulties associated with making sense of the idea of something non-physical (the mind) causing something to happen in the *physical* world. Usually event A can cause event B only if there is some causal *law* which connects events of type A with events of type B; for example, dropping the vase on the floor causes it to shatter *because* there are laws which connect physical impact with breaking. But in the case of mental causation, the cause itself is a *mental* event, and the dualist claims that mental substance is not subject to the causal laws that govern the physical world, because it is not *part* of the physical world. How can a mental event cause a physical event, if there are no causal laws connecting events of these types?

One popular interactionist response to this criticism is to say that it assumes too much about what causation can and can't be. If Hume is right, and what we observe in causation is nothing more than a 'constant conjunction' or regular connection between events of one kind and events of another, then surely we have a right to talk of mental events causing physical events, since we can directly observe a correlation between

events in our own mind and events in the physical world – for example when I decide to do something, and then do it! Beyond this idea of causation as a constant conjunction, the interactionist might say, we don't know anything about what causation *is*, so why shouldn't we believe that the kind of causation involved in mind-body interaction is possible?

A second conceptual problem has to do with location: immaterial minds are usually conceived of as lacking a physical location (after all, they are not part of the physical world). But usually causes must be in the same place as their effects – we do not usually expect to find causal **action at a distance**. It does not seem easy to understand how an unlocated thing could cause an event located at a specific place, without actually being at that place. This is the objection raised by Princess Elisabeth of Bohemia, in a letter written to Descartes in May 6th 1643:

> 'I have overcome my inhibitions and come right out with the question I put to the Professor, namely: Given that the soul of a human being is only a thinking substance, how can it affect the bodily spirits, in order to bring about voluntary actions? The question arises because it seems that how a thing moves depends solely on (i) how much it is pushed, (ii) the manner in which it is pushed, or (iii) the surface-texture and shape of the thing that pushes it. The first two of those require contact between the two things, and the third requires that the causally active thing be extended. Your notion of the soul entirely excludes extension, and it appears to me that an immaterial thing can't possibly touch anything else.'

Elisabeth notices – correctly – that Descartes' account of the mind (which she calls the 'soul') means that the mind must lack a physical location, but then points out that something can cause something else to move only by *touching it*, and for the mind to touch something it must first be located at the right place. Thus it seems impossible for the mind to affect the 'bodily spirits' which were believed in Descartes' time to control the actions of the rest of the body. Her point can be restated in entirely modern terms: philosophers often appeal to a principle of 'no action at a distance' when talking about causation, and this requires that anything that causes something else to happen must be located at the same place as the thing it affects causally; if the mind has no location then it cannot have any causal interaction with things that are located, such as the nerve cells that transmit electrical impulses to get our muscles to move.

Descartes replied to Elisabeth fifteen days later. His main point is that it is a mistake to try to understand 'the soul's power to move the body and the body's power to act on the soul in causing its sensations and passions' in terms of the very different 'notion' of how physical objects interact with each other. The notion of how the soul acts on the body, and *vice versa*, is for Descartes one of the 'basic notions that are like templates on the pattern of which we form all our other knowledge', and it would be wrong to try to explain such a 'basic notion' in terms of other ideas:

> 'When we try to explain some difficulty by means of a notion that isn't right for it, we are bound to go wrong; just as we are when we try to explain or define one of these notions in terms of another, because each of them is basic and thus can be understood only through itself. ... For example, when we try to use the imagination to conceive the nature of the soul, or when we try to conceive how the soul moves the body in terms of how a body moves a body. In the Meditations, which you were good enough to read, I tried to make conceivable the notions that are right for the soul alone, distinguishing them from the ones that are right for the body alone … Thus, I think we have until now confused the

notion of the soul's power to act on the body with the body's power to act on other bodies.'

Moreover, Descartes gives an example of how something can be thought to act on a physical 'body' without in any sense touching it, by talking about a theory of weight, which he himself rejected, but which (he claims) gives an example of the *kind* of action we suppose when the mind affects the body:

> 'Take for example what happens when we suppose that weight is a 'real quality' about which we know nothing except that it has the power to move the body that has it toward the centre of the earth. How do we think that the weight of a rock moves the rock downwards? We don't think that this happens through a real contact of one surface against another as though the weight was a hand pushing the rock downwards! But we have no difficulty in conceiving how it moves the body, nor how the weight and the rock are connected, because we find from our own inner experience that we already have a notion that provides just such a connection … I believe that this notion was given to us for conceiving how the soul moves the body.'

Again, the point can be restated in modern terms: gravity is an example of one thing causally affecting another without touching it (for example when the gravitational pull of the moon causes the tide to go in and out), and we have no difficulty understanding how that is possible; why then should the mind's 'acting on' the body require that the mind actually be *touching* the body. Better still, some interpretations of quantum mechanics explicitly *require* that some kind of 'action at a distance' can take place even when the two objects involved are too far apart for information to travel from one to the other in time.

However, it is not clear that Descartes' response is successful. First, it is hard to see why both physical interaction and mind-body interaction should not *both* be explained under one 'basic' or 'fundamental' idea of causation; moreover, examples of one located thing affecting another located thing 'at a distance' only go so far. What we want is a clear example of an *unlocated* thing affecting a located thing, and so far no uncontroversial example has been given. Even in Descartes' example of weight, the weight of the rock could plausibly be claimed to be located where the rock was – it is not as though weight is an *unlocated* feature that somehow affects the rock without being located at the same place as the rock.

As well as the **conceptual** issues for interactionism, there are **empirical issues**. Perhaps the most important is that it contradicts the widely-held view known as the **causal closure of the physical**: the view that every physical event has a sufficient physical cause. There is no *room* for mental events to cause physical events, since any physical event you care to name *already* has a cause in the physical world. A similar criticism can be made in terms of the **law of conservation of energy**, which states that the total amount of energy in a closed system cannot change: energy may be changed from one form to another (for example, from chemical to kinetic, when a bomb is detonated), but energy cannot be created or lost. If physical events in your brain or nervous system were being controlled by an unlocated, non-physical mind, that would mean that energy was coming *in* to the physical world from elsewhere, and so the physical world would no longer obey the law of conservation of energy.

A response is possible: the law of conservation of energy is intended to apply only to 'closed' systems: systems which do not interact with, or receive energy from, anything outside them. But the whole *point* of interactionism is that the physical world *does* interact with something non-physical, namely your mind; so the interactionist should never have expected that the physical world would obey the law of conservation of

energy. This criticism works only if we think that the physical world is a closed system – but that is exactly what the interactionist doesn't think.

A second response questions the status of laws such as the causal closure of the physical and the law of conservation of energy. Are they really empirically justified? Note first that, since they are universal claims (they are supposed to apply to everything everywhere), they cannot be *proved* through experiment, since that would require you to experiment on everything everywhere. Instead, scientists talk about laws such as these being 'well confirmed' by experiments, *i.e.* they are consistent with all our empirical observations *so far*. But of course, that doesn't guarantee that they are right – many theories of physics have been consistent with our observations for centuries and yet turned out in the end to be false (Newtonian mechanics is a notorious example of a 'useful but false' theory – that's why it's still taught in school physics lessons).

Alternatively, the interactionist might point to the fact that – on some interpretations of quantum mechanics – not every event in the world is causally determined by previous physical events; at the microphysical 'quantum level' events are apparently random or 'probabilistic' rather than determined. Thus some philosophers and neuroscientists have suggested that the immaterial mind might be able to control brain processes by exerting 'influence' over the behaviour of neurones within the cerebral cortex; because of the structure of the 'neural network' this influence could be magnified to amount to the mind controlling the brain. Nevertheless, this suggestion has not been widely accepted, since it is **hostage to fortune** in relying on the hope that the correct interpretation of quantum mechanics will be one that leaves room for genuine indeterminacy at the quantum level. Moreover, it does not deal with the question of why *my* mind has a causal influence only on *my* brain, and not on anyone else's, and indeed – when quantum behaviour is so widespread – why we find minds influencing only brains, and not other pieces of matter.

Week 5: Epiphenomenalism; the Problem of Other Minds

Last week, we looked at the issues facing **interactionist substance dualism**, according to which the mind and brain are distinct substances, but nevertheless interact causally. One way of avoiding these issues while remaining a dualist is to endorse **epiphenomenalism**. This is the view that our minds, though they exist, never actually cause any physical events, since the real causes of our actions are physical events in the brain. Although physical events can cause events in our mind that we are conscious of, and although it *seems* to us like we are making things happen 'out there' in the physical world, this is just an illusion – we are passengers in, rather than drivers of, our physical body, like a child with a toy steering wheel in the back seat of a car, who believes that he is actually controlling the vehicle. This view faces several difficulties: one noticeable one is that, if there is a problem explaining how non-physical minds can affect physical brains, there is no less of an issue for the epiphenomenalist to explain how physical brains can affect non-physical minds. The interaction problem is an *inter*action problem, and rejecting causal action in one direction (mind to brain) does not make it any easier to explain causal action in the other (brain to mind). Here are some other issues for epiphenomenalist dualism:

First, there is the allegation that, if epiphenomenalism is true, then we have to accept the **causal redundancy of the mental** – the claim that mental events, states, and properties have no causal role whatsoever. Here's how David Chalmers puts it (*n.b.* he calls epiphenomenalist dualism 'type-E dualism'); you may find it helpful to number each of his three objections in the margin to keep track of them:

> 'If type-E dualism is correct, then phenomenal states have no effect on our actions, physically construed. For example, a sensation of pain will play no causal role in my hand's moving away from a flame; my experience of decision will play no causal role in my moving to a new country; and a sensation of red will play no causal role in my producing the utterance 'I am experiencing red now.' These consequences are often held to be obviously false, or at least unacceptable.
>
> Still, the type-E dualist can reply that there is no direct evidence that contradicts their view. Our evidence reveals only regular connections between phenomenal states and actions, so that certain sorts of experiences are typically followed by certain sorts of actions. Being exposed to this sort of constant conjunction produces a strong belief in a causal connection (as Hume pointed out in another context); but it is nevertheless compatible with the absence of a causal connection. Indeed, it seems that if epiphenomenalism were true, we would have exactly the same evidence, and be led to believe that consciousness has a causal role for much the same reasons. So if epiphenomenalism is otherwise coherent and acceptable, it seems that these considerations do not provide strong reasons to reject it.
>
> Another objection holds that if consciousness is epiphenomenal, it could not have evolved by natural selection. The type-E dualist has a straightforward reply, however. On the type-E view, there are fundamental psychophysical laws associating physical and phenomenal properties. If evolution selects appropriate physical properties (perhaps involving physical or informational configurations in the brain), then the psychophysical laws will ensure that phenomenal properties are instantiated, too. If the laws have the right form, one can even expect that as more complex physical systems are selected, more complex states of consciousness will evolve. In this way, physical evolution will carry the evolution of consciousness along with it as a sort of byproduct.

> Perhaps the most interesting objections to epiphenomenalism focus on the relation between consciousness and representations of consciousness. It is certainly at least strange to suggest that consciousness plays no causal role in my utterances of 'I am conscious'. Some have suggested more strongly that this rules out any knowledge of consciousness. It is often held that if a belief about X is to qualify as knowledge, the belief must be caused in some fashion by X. But if consciousness does not effect physical states, and if beliefs are physically constituted, then consciousness cannot cause beliefs. And even if beliefs are not physically constituted, it is not clear how epiphenomenalism can accommodate a causal connection between consciousness and belief.'
>
> Chalmers, *Consciousness and its Place in Nature*

A second objection to epiphenomenalism comes from the **argument from introspection**: this is the claim that we are directly aware (through our faculty of 'introspection') of our mental events causing not only other mental events, but also of mental events causing physical events: I am directly aware not only that my conscious experience of pain causes a lowering of my mood (mental event causing mental event), but also of my conscious experience of pain causing me to move my hand away from the flame, or cry out, or seek medical attention. In each case it seems we have a direct introspective awareness of some of the contents of our mind being causally effective; but that is precisely what the epiphenomenalist denies.

The response to this on behalf of the epiphenomenalist is that we cannot infer facts about what causes what from our experience (even our 'inner' experience), because all we are aware of through experience – even when we observe cases of genuine causation – is one thing happening, followed by another thing happening. This point was famously made by Hume in his attempt to explain the source of our concept of causation. Applied here, the epiphenomenalist can say that *what* we experience through introspection is just a mental event, followed by a physical event; since what we experience would be the same *whether or not* the mental event really was the cause of the physical event, it cannot be the case that our inner 'experience' justifies the belief that mental events can be causally effective, since the best account of our inner experience (our 'phenomenology') is equally consistent with both epiphenomenalism *and* interactionism.

A third source of problems for epiphenomenalism is **issues relating to free will and responsibility**. This is a vast topic, but there are (broadly speaking) two ways to develop it. First, if you follow Descartes and take the view that you yourself, as a person, are just your conscious *mind,* and it turns out (as the epiphenomenalist claims) that your mind does not cause anything to happen in the physical world (including the movements of your own body), then *you* do not cause anything to happen in the physical world, and so cannot be held morally responsible (or accountable in any way) for anything that happens in the physical world. This would be like a defendant in court arguing that they were not responsible for a crime on the grounds that 'I didn't do it – my brain did'.

Alternatively, if you accept that you are both your conscious mind *and* your brain, you might worry that your brain – which is controlling your actions in the physical world – is in turn controlled by the strict, exceptionless laws of physics which control *everything* in the physical world, and so everything you do is simply the consequence of the circumstances you are in and the laws of physics. (This view, that human action is determined by the laws of nature plus the previous state of the universe, is known as **determinism**, and philosophers who claim that determinism rules out free will and moral responsibility are known as **hard**

determinists.) So in any given situation, there is only *one* possible way you can act: given the circumstances, you 'couldn't have done otherwise'. That sounds like a very powerful reason for thinking that you are *not* morally responsible for your actions, since many people believe that you should be held morally responsible for an action only if you could have refrained from doing it.

A response to this criticism may come in the form of developing one or another version of **compatibilism**, which is the view that freedom (and responsibility) are in fact compatible: the fact that my actions are governed by strict laws of nature does *not* prevent me from being 'free' and 'in control' of my actions. One very influential account was developed by David Hume: according to him, people have free will if they are capable of acting in accordance with their choices and decisions: as he put it, 'acting in accordance with the determinations of the will'. Hume claims that it does not *matter* that these decisions may have been caused in turn by some other factors which are beyond our control, or that the operation of the laws of nature on our physical brain guarantees that there is only one possible decision we can make in each situation; what matters is simply that our actions are 'in accordance with' our decisions. However, there is room for doubt about whether the epiphenomenalist can help herself to Hume's version of compatibilism here. Surely our 'decisions' and the 'determinations of our will' are *mental* events – and according to the epiphenomenalist, mental events never cause anything. So it *can't* be the case that our actions are consequences of our decisions. We might *think* that we are acting in accordance with what we consciously decide to do, but for the epiphenomenalist the causal story is the other way round: if anything, what we think of as a 'decision' is a *consequence* of the processes going on in our brains that control our actions, not a *cause* of those actions. (A very famous experiment – the **Libet experiment** – uses brain imaging technology to show (allegedly) that some of our decisions are *already* settled in our brains several seconds before we are consciously aware of having 'made up our minds.'

Dualism, the 'Problem of Other Minds', and Solipsism

Dualism makes one problem especially vivid: how do we know that other people *have* minds? If mental states are **reducible** to physical states of the brain, then this problem doesn't arise, since the existence of a correctly functioning brain is sufficient for the existence of a mind associated with it. But if the mind is a radically different kind of thing from the brain, as the substance dualist claims, then it is at least a coherent hypothesis that you are the only person who *has* a mind; everyone else is an automaton whose actions are controlled by their brains *without* any conscious mental states. (We will return to this possibility later on in the section on philosophical **zombies**.) Worse still, if you are an epiphenomenalist, then the conscious mind has no causal influence on behaviour whatsoever: there is no way you can infer the existence of mental states in other people on the basis of their observable behaviour because, according to epiphenomenalism, no observable behaviour is caused by mental states. These worries could lead you to adopt **scepticism about other minds** (the view that it is impossible to know that minds exist apart from your own), or worse still s**olipsism** (the view that *nothing* exists beyond you and your own experiences). Thus it can be seen that **some forms of dualism make it impossible to know other minds**. How could a dualist respond to these sceptical challenges? Here are some suggestions:

- John Stuart Mill developed a version of the **argument from analogy**: since we observe a correlation between mental states and outward behaviour in our own case, we are entitled (by 'analogy') to

conclude that similar outward behaviour in other people is caused by similar mental states. Here's how Mill makes his point:

> 'I conclude that other human beings have feelings like me, because, first, they have bodies like me, which I know, in my own case, to be the antecedent condition of feelings ; and because, secondly, they exhibit the acts, and other outward signs, which in my own case I know by experience to be caused by feelings. I am conscious in myself of a series of facts connected by an uniform sequence, of which the beginning is modifications of my body, the middle is feelings, the end is outward demeanour. In the case of other human beings I have the evidence of my senses for the first and last links of the series, but not for the intermediate link. I find, however, that the sequence between the first and last is as regular and constant in those other cases as it is in mine… by supposing the link to be of the same nature as in the case of which I have experience, and which is in all other respects similar, I bring other human beings, as phenomena, under the same generalizations which I know by experience to be the true theory of my own existence. And in doing so I conform to the legitimate rules of experimental enquiry. The process is exactly parallel to that by which Newton proved that the force which keeps the planets in their orbits is identical with that by which an apple falls to the ground. It was not incumbent on Newton to prove the impossibility of its being any other force ; he was thought to have made out his point when he had simply shown, that no other force need be supposed.'
>
> Mill, *An Examination of Sir William Hamilton's Philosophy* (1865), chapter XII

Objectors to this argument point out that what we observe is a correlation between mental states and behaviour in just *one* case (ourselves); how could this justify us in believing that every other case of human behaviour follows the same pattern? After all, observing just *one* white swan does not give us any right to conclude that all swans are white. One way of defending against this criticism is to claim that we are in fact aware of *many* instances of physical events having mental consequences and *vice versa*: almost everything that happens to us results in some kind of conscious mental experience, so we have observed a wide variety of instances of the different ways in which different kinds of physical event can cause different kinds of mental events. Nevertheless, the sceptic will be unconvinced: although we have observed many instances of mental event, we have done so in relation to just *one* mind. Compare: I cannot draw a valid conclusion about what the weather is like on earth *in general* by observing the weather in just one location on the globe, no matter how many different occasions I observe the weather on, and no matter how many different kinds of weather I experience at that location.

- Alternatively, we might think of belief in the existence of other minds as a case of **inference to the best explanation**: the theory that other people have conscious experience, and are motivated by beliefs and desires, in roughly the same way we are, is something that enables us to explain and predict their behaviour, and it is the theory's success in explaining observed behaviour that gives us a right to believe that it is true. Nevertheless, the **epiphenomenalist** cannot help herself to this response: like the eliminative materialist she believes that the *real* explanations of action are to do with the neurological structure of the brain, not our conscious choices and decisions – if the behaviour of others can be explained in principle solely by events within their brain, why do we need to treat them as having *minds* as well?
- It might be claimed that outward behaviour forms part of the *meaning* of terms which we use to ascribe mental states to others, such as '...is in pain' or '...believes that...' We are justified in applying such terms to other people because they exhibit the appropriate behaviour; but of course to ascribe mental states in this way is to treat other people as having minds. We are justified in treating other people as having minds because we are justified in ascribing mental states to them, and we are justified in ascribing mental states to them because they exhibit behaviour which is part of the *meaning* of the words involved. This approach is supported by thought about **how we learn to self-ascribe** mental states: we learn to ascribe mental states to ourselves because we hear other people ascribing them to us on the basis of our behaviour; thus – even in our own case – mental states are ascribed at least partly on the basis of behaviour rather than introspection.
- Wittgenstein is often interpreted as using the **private language argument** to show that doubt about the existence of other minds (or at least, other language users) is self-refuting. The central thought here is that the mere fact that we can use a meaningful language to formulate the possibility that other minds do not exist shows that there must be other language users, as a meaningful language must be one that is *shared* within a community, with recognized rules for correct and incorrect usage, and not just a 'private' language in which the words get their meaning by being correlated with my own private sensations.

Thus the private language argument is the argument **that there can be no such thing as a private language**, where a private language is one where the meanings of the words are fixed by the decision of an individual and not by the agreement of a community. Wittgenstein's reasoning is famously hard to follow, but the central point seems to be this: in a genuinely 'private' language there would be no difference between 'correct' and 'incorrect' uses of a word, as there would be no difference between the speaker actually using a word correctly and merely thinking he was using the word correctly; but this would be a 'language' with no rules – and such a language isn't really a language, just a sequence of meaningless noises. So the fact that I am *using* a language meaningfully is enough to show me that I am part of a community of language-users, and so there must be other minds, and solipsism must be false.

Exercise: how much of this excerpt from Wittgenstein's *Philosophical Investigations* can you understand? Underline/highlight any sentences which seem to express parts of the private language argument.

258. Let us imagine the following case. I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign "S" and write this sign in a calendar for every day on which I have the sensation.——I will remark first of all that a definition of the sign cannot be formulated.—But still I can give myself a kind of ostensive definition.—How? Can I point to the sensation? Not in the ordinary sense. But I speak, or write the sign down, and at the same time I concentrate my attention on the sensation—and so, as it were, point to it inwardly.—But what is this ceremony for? for that is all it seems to be! A definition surely serves to establish the meaning of a sign.—Well, that is done precisely by the concentrating of my attention; for in this way I impress on myself the connexion between the sign and the sensation.—But "I impress it on myself" can only mean: this process brings it about that I remember the connexion right in the future. But in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about 'right'.

259. Are the rules of the private language impressions of rules?— The balance on which impressions are weighed is not the impression of a balance.

260. "Well, I believe that this is the sensation S again."—Perhaps you believe that you believe it! Then did the man who made the entry in the calendar make a note of nothing whatever?—Don't consider it a matter of course that a person is making a note of something when he makes a mark—say in a calendar. For a note has a function, and this "S" so far has none. (One can talk to oneself.—If a person speaks when no one else is present, does that mean he is speaking to himself?)

261. What reason have we for calling "S" the sign for a sensation? For "sensation" is a word of our common language, not of one intelligible to me alone. So the use of this word stands in need of a justification which everybody understands.—And it would not help either to say that it need not be a sensation; that when he writes "S", he has something—and that is all that can be said. "Has" and "something" also belong to our common language.—So in the end when one is doing philosophy one gets to the point where one would like just to emit an inarticulate sound.—But such a sound is an expression only as it occurs in a particular language-game, which should now be described. 262. It might be said: if you have given yourself a private definition of a word, then you must inwardly

undertake to use the word in such-and-such a way. And how do you undertake that? Is it to be assumed that you invent the technique of using the word; or that you found it ready-made? 263. "But I can (inwardly) undertake to call THIS 'pain' in the future."—"But is it certain that you have undertaken it? Are you sure that it was enough for this purpose to concentrate your attention on your feeling?"—A queer question.—

264. "Once you know what the word stands for, you understand it, you know its whole use."

Week 6: Reductive accounts of the mind (I) – Logical behaviourism

To understand the debate fully over the next few weeks you need to ensure that you understand some **essential distinctions**. Fill in the gaps below! (You may want to work on rough paper first and then complete this pages as we go over in class) **:**

- We have distinguished **dualism** from **materialism**. Materialists claim that the world contains only _____

  _____

  They also claim that minds and mental properties are not _____

  _____

- Earlier we distinguished **reductive** from **eliminative** materialist theories of mind. While an eliminativist claims that statements about 'minds' should be rejected as part of a radically misleading 'folk psychology', a reductionist says that statements about minds can in principle be _____

  _____

- You already know the difference between **qualitative** and **numerical** or ('strict') **identity**. If *a* and *b* are qualitatively identical, they _____

  _____,

whereas if *a* and *b* numerically identical, they _____

  _____

- Last years' work introduced you to the distinction between **a priori** statements, which_____

  _____,

and **a posteriori** statements, which _____

  _____

- You also know that a **necessary** truth is one which_____

  _____,

while a **contingent** truth is one which _____

  _____

- Finally, you should remind yourself that the **verification principle** says that_____

_____

_____

Logical behaviourism

Philosophers distinguish **psychological behaviourism** from **logical behaviourism** (also known as '**analytical behaviourism**'). Psychological behaviourism is a view about the methods and aims of psychology: in short, that psychology aims to predict and control the behaviour of both human and non-human animals purely in terms of responses to stimuli and patterns of behaviour learnt through reinforcement. This kind of behaviourist 'recognizes no dividing line between man and brute' (as John D. Watson put it in 1913), since even human behaviour is to be explained in terms of learnt responses to stimuli rather than in terms of rational decision-making and conscious mental states. The classic example of a behaviourist experiment is known as **Pavlov's dogs –** a demonstration that reflex actions could be *conditioned* as well as innate. (Ivan Pavlov discovered that dogs could be conditioned to salivate when a bell was rung, even if no food was present – for a more detailed explanation see http://www.nobelprize.org/educational/medicine/pavlov/readmore.html)

Logical or 'analytical' behaviourism – which is the theory relevant to philosophy of mind – is

> 'a theory within philosophy about the meaning or semantics of mental terms or concepts. It says that the very idea of a mental state or condition is the idea of a behavioral disposition or family of behavioral tendencies, evident in how a person behaves in one situation rather than another. When we attribute a belief, for example, to someone, we are not saying that he or she is in a particular internal state or condition. Instead, we are characterizing the person in terms of what he or she might do in particular situations or environmental interactions.' (Stanford EoP, *Behaviourism*)

Note especially the claim at the start of that explanation: logical behaviourism is a thesis about what mental terms or concepts *mean*. Thus it can be defined as the view that **all statements about mental states can be analytically reduced without loss of meaning to statements about behaviour**.

Logical behaviourists (henceforth, just 'behaviourists') were motivated by two apparent advantages of their theory: one is that it enables us to say how statements about someone's mental states can be *true* without having to adopt a dualist theory of mind, i.e. behaviourism is compatible with **materialism**. The other apparent advantage is that behaviourism is compatible with the **verification principle** – since statements about mental states are explained in terms of actual or possible behaviour, such statements are in principle *verifiable*. But although behaviourism was very popular among both philosophers and scientists for a short period in the 20th century, it is now quite an unpopular view. (This is, of course, no reason for you to reject it – take it on its merits!) The main factors leading to its rejection were:

- The abandonment of the verification principle as a test for meaningfulness, with the result that people accepted that claims about mental states could be meaningful even if they are *not* explained in terms of observable behaviour.

- **Issues defining mental states satisfactorily**. These take two forms. First, there is the worry that it is not possible to specify *the* behaviour associated with any particular mental state: for example, your belief that it is raining (a mental state) could be revealed in indefinitely many actions depending on what other beliefs or desires you have – whether you want to get wet or stay dry, whether you believe than an umbrella is close at hand, whether you believe in an angry rain god who has to be propitiated by means of human sacrifice, and so on. (See the excerpt from Tim Crane below.) This suggests that the analysis of any particular mental state will need to be infinitely complex, as it will need to specify

the different behaviours associated with that mental state against the background of every other *possible* combination of mental states a person might have. But logical behaviourism is supposed to be a thesis about the *meaning* of our talk about mental states; it had better not be the case that the 'meaning' associated with each individual mental state is infinitely complex, since human minds are presumably not capable of understanding such complex meanings.

- Following on from this is a worry the about **circularity** of definitions of mental states in terms of behaviour– that the behavioural descriptions used to 'analyse' mental states actually presuppose an understanding of mental states, and it is no good to explain something in a way which can be understood only by someone who already understands it! One way to make this accusation is to claim that descriptions of behaviour will need to presuppose an understanding of the difference between doing something intentionally and merely moving in the same way by accident: thus a the difference between, say, the behaviours of 'falling down the stairs' and 'throwing oneself down the stairs' can be recognized only by someone who *already* has an understanding of the mental state of *intending* to do something. If our descriptions of behaviour *presuppose* an understanding of the mind, then they can't be used to *explain* the mind. To avoid this point, logical behaviourists often insisted that the descriptions of behaviour be purely 'scientific', expressible in physical terms which do not presuppose this kind of understanding of mentality – for example, 'Paul's hand moving upwards' rather than 'Paul saluting'.

  There is room for some doubt whether descriptions of behaviour *can* be given in purely physical terms, but even if it can, there is *another* way of making the circularity objection, which is this: each behaviourist analysis of a mental state will tell you what behaviour to look out for, given the *other* mental states you know the person to have. So you *already* need to know what other mental states a person has *before* you can start assigning mental states to them on the basis of their behaviour. This suggests that the behaviourist can't define any one mental state unless he has already defined all the *other* mental states associated with it. That makes the behaviourist project look like it can never even get started.

- Another issue is the **conceivability of mental states without associated behaviour**, or mental states which are accompanied by *no behaviour whatsoever* (see Putnam's 'Super Spartans'): examples like that suggest that behaviour is a sign or symptom of a mental state, not the same as the mental state, since we can imagine mental states which are never revealed in behaviour.

- Finally, there is the **asymmetry between self-knowledge and knowledge of other people's mental states**. As noted in week 1, we seem to have (at least some of the time) a kind of **privileged access** to our own mental states: we sometimes know the contents of our own minds *before* we act on them or reveal them in behaviour. Behaviourists struggle to say how this knowledge by **introspection** can be possible, since according to them, to be in a mental state just is to engage in the associated behaviour: it should not be possible to 'find' our own mental states in advance of behaving in a certain way, and really what the behaviourist ought to expect is that we can know our *own* minds only by examining our own behaviour or dispositions to behaviour. (One of the better jokes in philosophy: two behaviourists meet in the street. One says to the other, 'You're feeling well today. How am *I* feeling?')

Read the following excerpts and answer the questions that follow.

Hempel, *The Logical Analysis of Psychology* (1949)

(read from p.167 no.2-p.170)

1) What do you think Hempel means by a 'test sentence'?
2) What is the main conclusion Hempel draws about 'psychological statements'?
3) Explain the objection to his proposal which Hempel considers in section 5.
4) How does Hempel respond to this objection?

Putnam, *Brains and Behaviour*

'I believe that pains are not clusters of responses, but they are (normally, in our experience to date). Moreover, although this is an empirical fact, it underlies the possibility of talking about pains in the particular way in which we do. However, it does not rule out in any way the possibility of worlds in which (owing to a difference in the environmental and hereditary conditions) pains are not responsible for the usual responses, or even are not responsible for any responses at all...

Imagine a community of 'super-spartans' or 'super-stoics' - a community in which the adults have the ability to successfully suppress *all* involuntary pain-behaviour. They may, on occasion, admit that they feel pain, but always in pleasant, well modulated voices – even if they are undergoing the agonies of the damned. They do *not* wince, scream, flinch, sob, grit their teeth, clench their fists, exhibit beads of sweat, or otherwise act like people in pain or people suppressing the unconditioned responses associated with pain. However, they do feel pain, and they dislike it (just as we do). They even admit that it takes a great effort of will to behave as they do. It is only that they have what they regard as important ideological reasons for behaving as they do, and they have, through years of training, learned to live up to their own exacting standards.

What about verbal reports [of pain]? Some behaviourists have taken these as the characteristic form of pain behaviour ... let us undertake the task of trying to imagine a world in which there are not even pain *reports*. I will call this world the 'X-world'. In the X-world we have to deal with 'super-super-spartans'. These have been superspartans for so long that they have begun to suppress even *talk* of pain... X-worlders do not even admit to *having* pains. They pretend not to know either the world or the phenomenon to which it refers. In short, if pains are 'logical constructs out of behaviour', then our X-worlders behave so as not to have pains! - Only, of course, they do have pains, and they know perfectly well that they have pains.

If this last fantasy is not, in some disguised way, self-contradictory, then logical behaviourism is simply a mistake.'

1) Why is it that the conceivability of 'Super-spartans' and 'Super-super-spartans' shows that logical behaviourism is a mistake?
2) How would you respond to Putnam on behalf of a behaviourist?

Crane, *The Mechanical Mind* (1995, pp.51-2)

> 'thoughts cannot be fully defined in terms of behaviour: other thoughts need to be mentioned too. Each time we try to associate one thought with one piece of behaviour, we discover that this association won't hold unless other mental states are in place. And trying to associate each of these other mental states with other pieces of behaviour leads to the same problems. Your individual thought may be associated with many different pieces of behaviour *depending on which other thoughts you have.*
>
> A simple example will sharpen the point. A man looks out of a window, goes to a closet, and takes an umbrella before leaving his house. What is he thinking? The obvious answer is that he thought that it was raining. But notice that, even if this is true, this thought would not lead him to take his umbrella unless he also wants to stay dry *and* he believes that taking his umbrella will help him stay dry *and* he believes that this object is his umbrella ... if he didn't have these (doubtless unconscious) thoughts, it would be quite mysterious why he should take his *umbrella* when he thought it was raining.'

Gilbert Ryle and 'soft behaviourism'

Reading: The Concept of Mind, Chapter 1.

Ryle himself tried to distance himself from logical behaviourism; nevertheless his view is standardly described as '**soft behaviourism**', in discussions of his work, as he tries to retain the core behaviourist insight that the mind is *not* some extra mysterious thing causing our behaviour (which he called **the ghost in the machine**), and therefore denies that 'there are mental states and processes enjoying one sort of existence, and bodily states and processes enjoying another'; moreover he does claim that there is a close conceptual connection between ascriptions of mental states and outwardly observable behaviour: in fact he is prepared to say that 'Overt intelligent performances are not clues to the workings of minds; they are those workings.'

To diagnose the kind of error involved in thinking of the mind in this way, Ryle introduces the idea of a **category mistake** (or **category error**): this is the logical mistake made by people who 'are perfectly competent to apply concepts, at least in the situations with which they are familiar, but are still liable in their abstract thinking to allocate those concepts to logical types to which they do not belong'. His most famous example of such a mistake is when a 'foreigner visiting Oxford or Cambridge for the first time is shown a number of colleges, libraries, playing fields, museums, scientific departments and administrative offices. He then asks "But where is the University?" … It has then to be explained to him that the University is not another collateral institution … The University is just the way in which all that he has already seen is organized.' A similar category error, he suggests, is the mistake of thinking that statements about mental states are really descriptions of the mechanisms of some unseen, 'secret' machine which is the hidden cause of our observable behaviour.

So Ryle agrees with the behaviourists that it is a mistake to think of the mind, and mental processes, as the independently existing, partially hidden causes of outward behaviour. At the same time his account tries to avoid some of the more controversial commitments of a standard logical behaviourist such as Hempel. For example, he does not claim that the 'behavioural' translations of mental vocabulary must be completed using

only the language of science – every-day descriptions of actions will do just as well – and he does not claim that claims about minds must be explained in terms of actual behaviour; instead, in using mental vocabulary we often ascribe **dispositions** to action.

One important point about soft behaviourism is that it is *not* the same as standard 'logical' or 'analytical' behaviourism, because Ryle rejects the view that statements about mental states can be analysed into purely behavioural statements without loss of meaning: on his dispositional account, many attributions of mental states concern dispositions which may not be fully revealed through observable behaviour, and he accepts that his behavioural descriptions will to some extent presuppose an understanding of the mental states he is trying to explain. To some extent, he is even willing to accept that some actions might be unobservable (running through an argument in your head, for example), and so he rejects the view that all mental descriptions can be given in terms of *observable* behaviour. Moreover, his insistence that the range of dispositions associated with any given mental state can be 'indefinitely long' might lead us to suspect that he is not really interested in giving an analysis of the *meaning* of statements about mental states at all.

Read this extract from *The Concept of Mind* and answer the questions that follow.

> 'When we describe glass as brittle, or sugar as soluble, we are using dispositional concepts, the logical force of which is this. The brittleness of glass does not consist in the fact that it is at a given moment actually being shivered. It may be brittle without ever being shivered. To say that it is brittle is to say that if it ever is, or ever had been, struck or strained, it would fly, or have flown, into fragments. To say that sugar is soluble is to say that it would dissolve, or would have dissolved, if immersed in water.
>
> A statement ascribing a dispositional property to a thing has much, though not everything, in common with a statement subsuming the thing under a law. To possess a dispositional property is not to be in a particular state, or to undergo a particular change; it is to be bound or liable to be in a particular state, or to undergo a particular change, when a particular condition is realised. The same is true about specifically human dispositions such as qualities of character. My being an habitual smoker does not entail that I am at this or that moment smoking; it is my permanent proneness to smoke when I am not eating, sleeping, lecturing or attending funerals, and have not quite recently been smoking.
>
> In discussing dispositions it is initially helpful to fasten on the simplest models, such as the brittleness of glass or the smoking habit of a man. For in describing these dispositions it is easy to unpack the hypothetical proposition implicitly conveyed in the ascription of the dispositional properties. To be brittle is just to be bound or likely to fly into fragments in such and such conditions; to be a smoker is just to be bound or likely to fill, light and draw on a pipe in such and such conditions. These are simple, single-track dispositions, the actualisations of which are nearly uniform. But the practice of considering such simple models of dispositions, though initially helpful, leads at a later stage to erroneous assumptions. There are many dispositions the actualisations of which can take a wide and perhaps unlimited variety of shapes; many disposition-concepts are determinable concepts. When an object is described as hard, we do not mean only that it would resist deformation; we mean also that it would, for example, give out a sharp sound if struck, that it would cause us pain if we came into sharp contact with it, that resilient objects would bounce off it, and so on indefinitely. If we wished to

unpack all that is conveyed in describing an animal as gregarious, we should similarly have to produce an infinite series of different hypothetical propositions.

Now the higher-grade dispositions of people with which this inquiry is largely concerned are, in general, not single-track dispositions, but dispositions the exercises of which are indefinitely heterogeneous. When Jane Austen wished to show the specific kind of pride which characterised the heroine of 'Pride and Prejudice', she had to represent her actions, words, thoughts and feelings in a thousand different situations. There is no one standard type of action or reaction such that Jane Austen could say 'My heroine's kind of pride was just the tendency to do this, whenever a situation of that sort arose'.

Epistemologists, among others, often fall into the trap of expecting dispositions to have uniform exercises. For instance, when they recognise that the verbs 'know' and 'believe' are ordinarily used dispositionally, they assume that there must therefore exist one-pattern intellectual processes in which these cognitive dispositions are actualised. Flouting the testimony of experience, they postulate that, for example, a man who believes that the earth is round must from time to time be going through some unique proceeding of cognising, judging, or internally re-asserting, with a feeling of confidence, 'The earth is round'. In fact, of course, people do not harp on statements in this way, and even if they did do so and even if we knew that they did, we still should not be satisfied that they believed that the earth was round, unless we also found them inferring, imagining, saying and dbing a great number of other things as well. If we found them inferring, imagining, saying and doing these other things, we should be satisfied that they believed the earth to be round, even if we had the best reasons for thinking that they never internally harped on the original statement at all. However often and stoutly a skater avers to us or to himself, that the ice will bear, he shows that he has his qualms, if he keeps to the edge of the pond, calls his children away from the middle, keeps his eye on the life-belts or continually speculates what would happen, if the ice broke.'

Ryle, *The Concept of Mind* (1949)

1) What is it, according to Ryle, to be in a dispositional state?
2) What does Ryle mean by saying that 'the higher-grade dispositions of people... are indefinitely heterogeneous'?
3) What dispositional account would Ryle offer of the mental states (a) being proud; (b) believing that the ice is thin?
4) Do you think that Ryle's 'soft' version of behaviourism successfully avoids the criticism raised by Crane?

Week 7: Reductive accounts (II) - Type-identity

Analytic vs. ontological reduction

Notice that the previous **reductive materialist** account (logical behaviourism) proposed an **analytic reduction**. This is where the theorist proposes a thesis about the *meaning* of statements about the mind: she says that what a claim like '*S* believes that *p*' means that '-----', where the blank is filled in with an *analysis* in terms that do not presuppose a prior understanding of mental terms like 'belief' or 'desire'. The analysis is supposed to show that statements about, for example beliefs, are *really about* something else (for example, they are about 'observable behaviour' – and that, of course, means the we no longer need to belief in the independent existence of beliefs. We thought we were talking about beliefs, but really we are talking about behaviour.

For that reason, someone who puts forward an **analytic reduction** is also committed to an **ontological reduction** (remember, 'ontological' means 'to do with what exists'). If talk about beliefs is really nothing more than talk about behaviour, then the word 'belief' does not pick out a different, special kind of thing that exists in the world: we might say 'beliefs are *nothing more than* certain kinds of behaviour'. An **ontological reduction** of this kind happens when a theorist claims that things of type *X* are really just things of type *Y*. So, for example, someone who claims that a mind is nothing more than a correctly functioning brain has suggested an ontological reduction of the mind to the brain.

Can you have a theory that proposes an **ontological reduction** without an **analytic reduction**? Yes. It's a strange feature of words and concepts in our language that two *different* words or concepts can pick out the same actually existing thing. So for example 'Peter Parker' and 'Spiderman' are two different names for the same person, and 'water' and 'H_20' arguably pick out two different *concepts* which nevertheless refer to the same thing in the world. For that reason, two words which mean different things (they pick out different concepts) can still turn out to refer to the same natural phenomenon. Think of the concepts of *lightning* and *cloud-based electrostatic charge*: they are different concepts (someone could understand one without understanding the other) but they are really talking about, or referring to, the same kind of thing in the world.

Exercise: make a diagram showing the relationship between words/phrases, concepts, and things in the world.

Types and tokens

How many words are written here?

THE THE

One answer is: one. There is one **type** of word here, which has two 'instances'. Another answer is, two. There are two **tokens** of the word 'the' written on the page. This helps us to understand the logical difference between **type-identity** and **token-identity**. Token-identity is when one particular thing is identical (numerically the same as) one other particular thing. You could think about Peter Parker and Spiderman again, or – for the classic philosophical example – the discovery that the 'evening star' Hesperus, and the 'morning star' Phosphorus are in fact one and the same thing – the planet Venus. To believe that Hesperus is Phosphorus is to commit yourself to the existence of *one* thing which has two different names.

A **type-identity** theorist, on the other hand, claims that *types* or *kinds* of mental state are identical with types or kinds of brain state: for example, that *pain* (a kind of mental state, which unfortunately has many different instances or 'tokens') is identical with the stimulation of C-fibres in the brain. Identity here is intended to be

understood as *numerical* identity – i.e. we discover that what we thought of as two different things are really *one* thing. To say that pain is C-fibre stimulation is to commit yourself to the view that this is *one* kind of phenomenon, which has two different names (and maybe two different concepts associated with it).

The main reason for adopting this approach (apart from the obvious one of avoiding **dualism**) is that it removes the need to give an account of the **causal relationship** between minds and brains. If pain just *is* a kind of brain state, then there is no need to explain how pain is *caused* by certain brain states: it is senseless to ask how one thing causes *itself* to occur.

An obvious objection to this suggestion is that we have different *beliefs* about pain and C-fibre stimulation: for a start, we believe that pain hurts, but we do not believe that C-fibre stimulation hurts. But this does not stop the two things being identical. Think of Clark Kent and Superman again: Lois Lane thinks of Superman in a completely different way from how she thinks of Clark Kent, but this does not stop Clark Kent and Superman being different names of the same one person. You might want to say that Loius Lane has two different *concepts* – one for each name – but as we said before, two different concepts can turn out to represent the same one thing in reality. (A completely tangential question: can she be in love with one of them and not with the other, even if they are the same person with the same personality?)

A second problem is that **brain states have precise spatial locations which thoughts lack**. Each individual brain state is located where the associated neurons are – my brain state of *having C-fibres firing* is located *at* my C-fibres, while it is hard to provide such a precise location for pains – and in any case, a pain in one's foot is felt *in the foot*, while the C-fibre firing stretches all the way up to the brain. However, this objection can be answered by appealing to the 'same thing, different concepts' approach explained earlier. There's nothing in the concept of 'Clark Kent' to suggest that he has superpowers; but he does! Concepts don't have to tell us the whole truth about the nature of something. Likewise, the concept of 'pain' might not reveal that the pain is located where the C-fibres are firing. Indeed, it's common for scientific investigation to reveal new facts about something which were not included in our original concept. It was not part of the original concept of water that water contains hydrogen and oxygen atoms in combination, and it would be foolish for someone to have argued that water isn't H_2O by saying 'we don't usually think of water containing hydrogen...'. Likewise, the fact that we don't usually *think* of mental states as having a precise location doesn't mean that we can't find out through neurophysiological research that they *do*.

A third problem for type-identity theorists is that, although in this world pain is accompanied by C-fibre stimulation in every brain we can observe, it could have been the case that pain was associated with a different kind of neurological event (for example, D-fibre stimulation); indeed, it could have been the case that pain existed among the silicon-based life-forms of science fiction, who have nothing identifiable as a brain. For that reason, type-identity theorists often endorse a claim of **contingent identity**: although pain is, in fact, identical with C-fibre stimulation, it could have been the case that pain was identical with something else instead; the claim 'pain is C-fibre stimulation' is contingently rather than necessarily true. Many philosophers have now come to doubt that contingent identity is a coherent idea: for how could it have been the case that one thing should fail to be the same as itself? That is the point made by the excerpt from Kripke below.

Smart, *Sensations and Brain Processes* (1959)

Read pp.144-150

1) Give an accurate statement of the content of the thesis 'that sensations are brain processes'
2) Summarize objections 1-3 and Smart's responses to them.

Kripke, *Identity and Necessity (1971)*

R*ead from p.* 158-end.

1) In what way, according to Kripke, is the identification between heat and the motion of molecules like that between pain and a certain kind of brain state?
2) In what ways are these identifications different?

A note on Kripke's argument

Kripke argues *against* mind-brain type identity theory, on the grounds that (i) type-identity theory requires that the identity between pain and a particular brain state must be *contingent* ('we can imagine the brain state existing, though there is no pain at all... one might imagine a creature being in pain, but not being in any specified brain state at all, maybe not having a brain at all. People even think they can imagine … totally disembodied creatures … we can imagine definite circumstances under which this relationship would have been false.'), but (ii) the identity in question could *only* be *necessary* as a matter of pure *logic*. Since he (like most people) takes point (i) to be obvious, he spends most of his time arguing for the second point.

He first claims that terms like 'pain' and 'brain state *X*' (and names in general) are what he calls **rigid designators** – they always pick out the *same* thing in every possible world we might happen to be talking about. From that it follows that, if pain *is identical with* brain state X, it must be identical with it in *every* possible world, since if 'pain' and 'brain state *X*' pick out one thing in this world, they must pick out the same one thing in every world, as long as they remain rigid designators.

Then he considers a potential counter-example. Physicists tell us that heat is the motion of molecules. On Kripke's account that identity should be necessary – it should hold in all possible worlds. But surely we can imagine possible worlds in which there is the motion of molecules without heat – so this is an example of contingent identity. Not so, says Kripke. The situation you are imagining is one in which there is still *heat*, but the creatures in that world lack the powers to experience it in the same *way* we do: just as there would still be such a thing as heat in this world even if every creature capable of sensing it died out, there would still be heat in a possible world in which creatures sensed temperatures differently, and even if the *feelings* the creatures had when exposed to heat and cold were exactly the reverse of ours, that wouldn't mean that 'heat was suddenly turned to cold'. As long as you have the right kind of molecular motion, you have heat – regardless of how creatures are set up to sense or experience it. So this is not an example of contingent identity.

Turning now to pain itself, Kripke points out that the way pain feels to us is not, like the way we sense heat, a contingent feature we use to identify something that could have felt differently to us: instead it is an 'essential property' of pain that it feels painful to the person who has it. So it makes no sense to take the same approach as in the case of heat, saying that pain is identical with a certain physical state and that this pain state could still have existed even in worlds where people sensed it differently and it was no longer painful: the idea of a world in which no-one feels anything as painful is a world in which there is no pain, whereas a world in which no-one senses heat could still be a world in which heat exists.

Here's the short version: the 'contingent identity' theorist wants to say that pain is a certain brain state, but might not have been (pain could have been a different brain state instead). Kripke says that this makes no sense: all we can say is that pain is a certain brain state, and that brain state might not have felt painful to us. But that is the same as saying that pain might not have felt painful to us, which is absurd. So it cannot be true that pain is identical with a brain state.

Week 8: Functionalism

Multiple realizability

A major reason why **type-identity** is a currently unfashionable theory of the mind is the problem of **multiple realizability**, first raised in the 1960s by Hilary **Putnam**:

> '[the type-identity theorist] has to specify a physical-chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of suitable physical-chemical structure ; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusa, and certainly feel pain), etc. At the same time, it must *not* be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain [certain with respect to the laws of nature] that it will also be a state of the brain of any extra-terrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may *be* pain. ... it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical 'correlate' of pain. But this is certainly an ambitious hypothesis.
>
> Finally, the hypothesis becomes still more ambitious when we realize that the brain state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say 'hungry'), but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly likely that we can do this.'
>
> Putnam, *Psychological Predicates* (1967)

Putnam was keen to stress that his argument is empirical: based on what we know about the world, isn't is overwhelmingly improbable that *every* creature in the universe that feels pain does so in virtue of *one* kind of brain state? Surely it's more likely that pain is 'realized' as one kind of brain state in humans, another in octopuses, and another in alien creatures we haven't met yet. There doesn't seem to be anything in the laws of nature that rules out pain being felt by silicone-based lifeforms whose 'brains' are made up of entirely different stuff from human brains.

If you're convinced by Kripke's rejection of **contingent identity**, and believe that any identity-claim between pain and a particular brain-state must hold 'in all possible worlds', then the problem of multiple realizability becomes stronger still: we can say that, even though pain in humans is correlated with brain state *A*, it *could have been the case* (in some other possible world) that pain was correlated with brain state *B* – and this would be enough to show that pain is not *identical* with brain state *A*, as pain is 'realizable' through either brain state.

The obvious response to this on behalf of the identity-theorist is to distinguish between 'pain for humans' and 'pain for octopuses': if the claim she is making is merely that 'pain for humans' is identical with brain state *A*, it doesn't matter that there are (or might be) other creatures whose 'pain' is correlated with different kinds of brain state. But that might not strike you as particularly plausible: telling us that 'pain for humans' is correlated with brain state *A*, and 'pain for octopuses' is correlated with brain state *B* does nothing to tell you what it is about states *A* and *B* that makes them both *pain* states; what is it that they have in common in virtue

of which they are both (kinds of) pain? We won't have an account of *what pain is* until we can say what it is that is the *same* in every creature that feels pain – and we won't do that by identifying different brain states for different species.

Moreover, it seems that there is strong evidence that mental states are 'multiply realizable' even among different members of the same species: a feature of the human brain is its **plasticity**, i.e. the feature that different parts of the brain are capable of adapting to perform different functions if required. For example, if one area of the brain is damaged by injury or disease, another area may take over the tasks normally performed by the damaged area. Because of this, we cannot be confident in identifying kinds of mental states with types of brain state even if the identification is limited to *human* brains: it may be that the mental state which is correlated with state *A* in one person's brain is in fact correlated with a completely different state in the brain of someone else. How could there, for example, be a specific configuration of neurones such that every single person who believes that Paris is the capital of France has precisely that configuaration of neurones in their brain?

Functionalism: roles and realizers

We can specify a **role** without thereby specifying which entity performs that role: we can talk about the role of *Hamlet* while leaving it open which actor actually performs as that Prince of Denmark – Laurence Olivier or David Tennant, say. So there is a distinction between a **role** and its **realizer**. Moreover, the role itself can often be specified in terms of a distinctive **function** – a bottle-opener functions to open bottles, and a mousetrap functions to trap mice. Indeed, it seems reasonable to say that anything that has the function of trapping mice just *is* a mousetrap, regardless of its shape, size, or the material it is constructed from. Specifying a role does not necessarily impose any limitations on the construction of the thing that realizes that role.

Exercise: what functional role would you associate with each of these?

- Oven _____
- Telephone _____
- Computer _____
- Shoe _____
- Diary _____
- Gun _____
- Table _____

So: some kinds of thing are identified not by their physical properties, but rather by the role they play. Functionalism takes this insight and runs with it, claiming that mental states are identified by the *role* they play, not by the physical properties of whichever brain state is *realizing* that role within one or another person's neural network. Here's how Jerry **Fodor** puts the point:

> 'The intuition underlying functionalism is that what determines the psychological type to which a mental particular belongs is the causal role of the particular in the mental life of the organism. Functional individuation is differentiation with respect to causal role. A headache, for example, is

identified with the type of mental state that among other things causes a disposition for taking aspirin in people who believe aspirin relieves a headache, causes a desire to rid oneself of the pain one is feeling, often causes someone who speaks English to say such things as "I have a headache" and is brought on by overwork, eyestrain and tension. This list is presumably not complete. More will be known about the nature of a headache as psychological and physiological research discovers more about its causal role.

Functionalism construes the concept of causal role in such a way that a mental state can be defined by its causal relations to other mental states.'

Fodor, *The Mind-Body Problem* (1981)

The central advantage of functionalism, then, is that it allows for mental states to be **multiply realized**: pain is the *same* mental state for me, an octopus, and an alien from Alpha Centauri, even though (owing to our radically different biological make-ups) there are no brain states we share in common. What makes my pain and your pain both instances of the *same* kind of mental state is simply that there is something realizing the same *role* in my mind as is realized by something else in the alien's mind. Sameness of mental state is sameness of causal role, not sameness of physical structure.

Functionalism is, for many philosophers, a development of (and alternative to) logical behaviourism: mental states are to be defined, not merely in terms of the behaviour they lead to (one kind of 'output'), but also in terms of their 'inputs' (what sort of things lead to, or cause, one to be in that mental state). Crucially, these causal relationships *include* relationships to *other* mental states. Since that's quite a hard idea to grasp, read this excerpt from Ned Block for another explanation of the view:

'One characterization of functionalism that is probably vague enough to be acceptable to most functionalists is: each type of mental state is a state consisting of a disposition to act in certain ways and to have certain mental states, given certain sensory inputs and certain mental states. So put, functionalism can be seen as a new incarnation of behaviorism. Behaviorism identifies mental states with dispositions to act in certain ways in certain input situations … Functionalism replaces behaviorism's "sensory inputs" with "sensory inputs and mental states"; and functionalism replaces behaviorism's "dispositions to act" with "dispositions to act and have certain mental states." Functionalists want to individuate mental states causally, and since mental states have mental causes and effects as well as sensory causes and behavioral effects, functionalists individuate mental states partly in terms of causal relations to other mental states.' Block, *Troubles with Functionalism*

Consequences: materialism, identity and reduction

Functionalists have been quick to point out that, although their theory is *compatible* with materialism – i.e. their account of the nature of mental states requires only the existence of physical matter – nevertheless it is not the case that functionalism *rules out* the possibility of minds being disembodied spirits or 'mental substances'. Functionalism is officially neutral on the subject of what minds are made of: minds can in fact be made of anything that is capable of the right kind of functional organization, and if there were immaterial ghosts they would be able to have minds, so long as whatever stuff they are made of is capable of realizing the right kind of functional organization. However, most functionalists believe that the human mind, at least, is constituted by the functional organization of the physical matter of the brain.

We have seen that functionalists reject **type-identity** theories, which identify a *type* of mental state (e.g. pain) with a *type* of physical state (e.g. the firing of C-fibres). This is because they want to allow that the brain-state which accompanies *my* pain might be a different type of brain state from the one that accompanies pain in octopuses or alien life forms; what our 'pains' have in common is sameness of the functional role they play. However, functionalists can and sometimes do endorse **token-identity**; this is the view that particular or 'token' mental states are identical to specific states in the brain: although pain in general is not identical to a *type* of mental state, any specific pain I have will be identical to some specific brain state I am in. In particular, my pain will be identical to whatever in my brain is fulfilling the functional role of a pain-state for me at the moment. This view is not a 'mainstream' one within functionalism, though: it is more common for functionalists to say that, strictly speaking, mental states such as pain should be identified with the functional *role* itself, not with the brain state that occupies or 'realizes' that role. While my (token) pain might be realized by a particular brain state at a particular time, it is not the case that my pain literally *is* that brain state.

Functionalists sometimes deny that their view is a **reductive** theory of the mind – i.e. functionalists do *not* believe that the 'higher-level' talk about mental states in terms of their function can be explained in terms of, or translated into, 'lower-level' talk about physical states of the brain. They draw an analogy between 'higher-level' and 'lower-level' sciences: the concepts and entities involved in higher-level sciences such as biology, geology and meteorology cannot be reduced to the concepts of fundamental physics; many of the explanations we need can only be given in the vocabulary of the 'higher level' science. In just the same way, the functionalist claims, the functional concepts we employ in describing our mental states may be irreducible – there is no way to translate the high-level theory into the language of 'low-level' fundamental physics without massive loss of explanatory power. Nevertheless, there is a sense in which some forms of functionalism *can* be described as reductive: if it is claimed that the nature of mental states can be described solely in terms of their causal relationships to one another and to their inputs and outputs, then it seems that some form of 'reduction' has been proposed, since we can now translate mental vocabulary into 'functional' vocabulary which does not presuppose an understanding of mental concepts. The functionalist has a response, of course: according to functionalism, a mental state can only be defined in terms of its relations *to other mental states* – so there is no prospect of 'reducing' talk about a given mental state in a way which does not presuppose an understanding of *other* mental states. (See 'Further Reading' for more on this question: Ned Block claims that functionalism *is* reductive, while Fodor is convinced that it isn't reductive.)

Perhaps the distinction between 'analytic' and 'ontological' reduction can help here. The A2 specification describes functionalism as the view that **all mental states can be reduced to functional roles which can be multiply realised**. This can't mean that we can **analyse** talk about mental states in terms that don't presuppose an understanding of mental states (as every definition of a mental state will need to presuppose an understanding of other mental states); however, there is some kind of **ontological reduction** here, as there is clearly a claim that being a mental state is 'nothing more than' being a particular functional role – although these 'functional roles' cannot be ontologically reduced further to be 'nothing more than' particular physical states or types of physical states, since the roles themselves can be multiply realized and are not associated with any particular arrangement of physical matter. For that reason, make sure you *never* claim that 'functionalists reduce the mental to the physical' – they don't!

Further reading:

Ned Block, *What is Functionalism?*

Jerry Fodor, *the Mind-Body problem*

Ned Block, *Troubles with Functionalism*

Eliot Sober, *Panglossian Functionalism and the Philosophy of Mind,* section 3

Daniel Dennett, *Brainstorms*

Week 9: Different kinds of functionalism

The mind as software in the brain

The classic example of something that can be 'multiply realized' is a computer program. The same set of instructions can be encoded in different programming languages, and followed by computers of radically different kinds: in fact, the earliest computers were mechanical, then electro-mechanical, then used vacuum tubes, then transistors, then silicone chips. Early computer programs were inputted on punched cards, and early computers could fill a whole room.

Here is a selection of what is functionally the *same* computer program, written to be followed by different computers running different programming languages. Can you tell what it does?

C++:
#include <iostream>
int main()
{
std::cout << "Hello, World.";
}
	Perl:
	print "Hello World\n"
	Visual basic:
	Sub Main()
	MessageBox("Hello World")
	End Sub

Linux shell script: #!/bin/sh echo "Hello World"	*Javascript:* document.writeln("Hello, World");

What all modern computers have in common is that they follow sets of instructions which tell them what to do next, based on what state they're in, and what input they are receiving. This idea of a computer as a rule-following machine was developed by Alan Turing in 1936, and people describe any machine which follows rules in the way he described as a '**Turing Machine**' in his honour. All modern computers run on essentially the same line as Turing machines. Some functionalists – **machine functionalists** – believe that the mind is, basically, a Turing machine, and that each mental state can be identified as a particular state which that machine might be in. What makes this a kind of functionalism is that it doesn't matter what kind of material is *realizing* the 'program' running on the Turing machine; what identifies a mental state is simply its relationship to inputs and outputs, and to other mental states. So the mental state *being in pain* could be described as the state which results from certain inputs (e.g. injury), gives rise to characteristic outputs (saying 'ouch!') and produces other mental states (e.g. loss of good humour).

Viewing: Horizon, *the Strange Life and Death of Dr Turing*

Read this excerpt by Ned Block and answer the questions that follow:

> 'A Turing-machine table lists a finite set of machine-table states, $S_1 . . . S_n$; inputs, $I_1 . . . I_m$; and outputs, $O_1 . . . O_p$. The table specifies a set of conditionals of the form: if the machine is in state S_i and receives input I_j, it emits output O_k and goes into state S_l. That is, given any state and input, the table specifies

an output and a next state. Any system with a set of inputs, outputs, and states related in the way specified by the table is described by the table and is a realization of the abstract automaton specified by the table.

To have the power for computing any recursive function, a Turing machine must be able to control its input in certain ways. In standard formulations, the output of a Turing machine is regarded as having two components. It prints a symbol on a tape, then moves the tape, thus bringing a new symbol into the view of the input reader. For the Turing machine to have full power, the tape muse be infinite in at least one direction and movable in both directions. If the machine has no control over the tape, it is a "finite transducer," a rather limited Turing machine. Finite transducers need not be regarded as having tape at all. Those who believe that machine functionalism is true must suppose that just what power automaton we are is a substantive empirical question. If we are "full power" Turing machines, the environment must constitute part of the tape.

Machine functionalists generally consider the machine in question as a probabilistic automaton - a machine whose table specifies conditionals of the following form: if the machine is in S_a, and receives I_b, it has a probability p_1 of emitting O_1; p_2 of emitting O_2; . . . p_k of emitting O_k; r_1 of going into S_1; r_2 of going into S_2; . . . r_n of going into S_n. For simplicity, I shall usually consider a deterministic version of the theory.

One very simple version of machine functionalism (Block & Fodor, 1972) states that each system having mental states is described by at least one Turing-machine table of a specifiable sort and that each type of mental state of the system is identical to one of the machine-table states. Consider, for example, the Turing machine described in the table (cf. Nelson, 1975):

	S_1	S_2
nickel input	Emit no output Go to S_2	Emit a Coke Go to S_1
dime input	Emit a Coke Stay in S_1	Emit a Coke & a nickel Go to S_1

One can get a crude picture of the simple version of machine functionalism by considering the claim that S_1 = dime-desire, and S_2 = nickel-desire. Of course, no functionalist would claim that a Coke machine desires anything. Rather, the simple version of machine functionalism described above makes an analogous claim with respect to a much more complex hypothetical machine table. Notice that machine functionalism specifies inputs and outputs explicitly, internal states implicitly (Putnam 1967, p. 434) says: "The S_i, to repeat, are specified only *implicitly* by the description, i.e., specified *only by* the set of transition probabilities given in the machine table"). To be described by this machine table, a device must accept nickels and dimes as inputs and dispense nickels and Cokes as outputs. But the states S_1 and S_2 can have virtually any natures (even nonphysical natures), so long as those natures

connect the states to each other and to the inputs and outputs specified in the machine table. All we are told about S_1 and S_2 are these relations; thus machine functionalism can be said to reduce mentality to input-output structures. This example should suggest the force of the functionalist argument against physicalism. Try to think of a first-order physical property that can be shared by all (and only) realizations of this machine table!

…

These notions of Turing Machines and machine tables were historically very important in the development of functionalism. This was for two reasons:

- they showed us how to simultaneously define a system of internal states, which interact with each other and with input and output. This answers the circularity worries that plagued the behaviorist.
- they showed us how to understand the internal states so defined in such a way that they can be implemented or realized via different physical mechanisms. Turing Machines are computationally equivalent to machine tables, and a machine table's states can be realized in a variety of ways.

Originally functionalists said that our minds were Turing Machines, and that mental states like belief and desire were just different states of this Turing Machine, in the same way that ZERO and FIVE and TEN are different states of our Coke Machine.

But remember that Turing Machines and machine tables are just two ways (mathematically elegant way) of formally specifying a piece of software. They have some distinctive features that other ways of specifying a piece of software do not. (This is akin to the differences between different programming languages.) So even if the analogy between minds and software is correct at a general level, the attempt to spell this analogy out in terms of Turing Machines and machine tables might face special problems. And indeed it does.

Machine tables are defined in such a way that they can be in only one state at a time. But typically we think of mental states as being states that a mind can be in several of, at the same time. For instance, one of my mental states is the belief that Harvard is in MA. Another mental state is the desire to finish writing up this web page. I am right now in both of these mental states. So we cannot identify mental states of this sort with states of a machine table.

We cannot make any sense of different machine tables having states in common. If your Coke machine does not implement exactly the same machine table as my Coke machine, then it is not possible for our Coke machines ever to be in the same machine state. We cannot say that both of our Coke machines are in state ZERO, for instance. State ZERO is defined in terms of the entire machine table it's part of. If our Coke machines implement different machine tables, then all of their machine states must therefore be different. Now, it is likely that the software my brain is running is somewhat different than the software your brain is running. When exposed to the same environmental stimulation, even from birth, people come to be in different mental states. So if we identified our mental states with states of a machine table, it would then be impossible for you and I to have any of

the same mental states. This is an absurd result. No doubt you and I differ in some of our mental states. But we also have many mental states in common.

Because of these difficulties, functionalists have moved away from Turing Machines and machine tables as models of the mind. Nowadays they spell out the analogy between minds and software in a slightly different way.' Block, *Troubles with Functionalism* (1980)

Questions:

1) Draw a Turing machine!
2) What do you understand as the difference between 'probabilistic' and 'deterministic' Turing machines?
3) Explain Block's two problems for machine functionalists.
4) Draw a probabilistic machine table showing how you would respond to being offered a doughnut by your friend, depending on what state you are already in.

Putnam's version of machine functionalism

In his 1967 article *Psychological Predicates*, Putnam argues for a version of machine functionalism. You should aim to read the article in full; to help you make sense of it, here is a summary of the key points:

- Putnam begins by going through some objections to type identity which he answers in ways which you are already familiar with: that we might believe and hence know different things about 'pain' and 'brain state *S*' (but this shows that the concepts are different, not that the concepts pick out different properties); that pain would have to have the same *location* as the brain state it is identical with (yet a mirror image is light reflected from an object, but it is in a different place from that object); that all we can say is that pain is 'correlated' with brain state *S* (but then we need to explain what the pain 'is', and *why* the correlation holds). This is Putnam clearing out *bad* objections before moving on to his own good ones.
- Putnam says that his reason for saying that pain is *not* a brain state is simply that 'another hypothesis is more plausible' - namely 'that pain is a functional state of a whole organism'.
- Putnam introduces the idea of a 'Probabilistic Automaton'. This is a Turing Machine (i.e. it changes state, and emits 'outputs' depending both on its inputs and what state it is already in), but it is 'probabilistic' in the sense discussed by Block: that its 'Machine Table' does not *determine* a particular response to each situation but merely assigns a probability to express the chance that you will take that response rather than any other. (This is to take account of the fact that humans do not seem *guaranteed* to respond in the same way in any given situation.)
- Putnam's thesis, then, is that all organisms capable of feeling pain are Probabilistic Automata, and that 'being capable of feeling pain *is* possessing an appropriate kind of Functional Organization' (Putnam says that Functional Organizations are described by 'machine tables' of the form the Turing Machine theorist proposes.) Further, actually *being* in pain, for Putnam, is the combination of having this appropriate kind of 'functional organization' and having certain kinds of 'sensory inputs'.
- Comparing his theory with the 'brain-state' theory of the type-identity theorist, Putnam notes that his view can, and type-identity cannot, deal with multiple realizability. (This is the passage quoted earlier.) He also notes that it is reasonable to expect other organisms which have similar mental states to have similar 'functional organizations': as he says, 'we would not count an animal as *thirsty* if its 'unsatiated' behaviour did not seem to be directed toward drinking and was not followed by 'satiation for liquid'.
- Comparing his view with logical behaviourism, Putnam notes that behaviourism is 'hopelessly vague' (it can't specify the 'relevant behavioural disposition' associated with being in pain as anything other than 'the disposition … to behave as if *X* were in pain'; secondly, functionalism allows that two behaviourally equivalent creatures may differ in respect of whether they are in pain – for example one person with all motor fibres cut (who cannot 'behave' in any way), as opposed to a person with both nerve fibres *and* motor fibres cut (who can neither feel or express pain). Finally, he makes the point that functionalism *explains* the behaviour-disposition associated with pain in terms of some underlying state, whereas behaviourism can't explain the behaviour in terms of the underlying pain-state, since the pain just *is* the behaviour.

Nonmachine functionalism and psychofunctionalism

Nonmachine functionalists say that mental states cannot be modelled by anything as simple as tables which give 'computational' rules connecting them with each other and their characteristic inputs and outputs. Instead they maintain that mental states are individuated by their **causal role** – i.e. which causal relationships they have with other mental states, inputs, and outputs; what causes them, and what they cause in turn. One advantage of this approach is that it allows that a person may be in more than one mental state at one time, since the mind is no longer explained on the model of a Turing machine that can only be in one state at a time. Another advantage is that the 'function' of a given mental state may be defined in more complex ways than simply producing a pre-determined output and/or moving on to another mental state, as the Turing machine does.

One particularly interesting version of nonmachine functionalism is known as **psychofunctionalism**, which says that mental states should be identified by whatever causal role they have in whatever turns out to be the correct scientific theory of psychology:

> 'Psychofunctionalism can be defined as the doctrine that mental states are constituted by causal relations among whatever psychological events, states, processes, and other entities – as well as inputs and outputs – actually obtain in us in whatever ways those entities are actually causally related to one another. Therefore, if current theories of psychological processes are correct in adverting to storage mechanisms, list searchers, item comparators, and so forth, Psychofunctionalism will identify mental states with causal structures that involve storage, comparing, and searching processes as well as inputs, outputs, and other mental states.'
>
> Block, *Troubles with Functionalism*

Week 10 – Qualia and Functionalism

You should remember that **qualia** are the 'felt qualities of experience'; they are the properties of mental states that are responsible for there being something 'it is like' to be in that state. Seeing that a tomato is red, for example, is not merely a *cognitive* achievement – not merely gaining the ability to use 'this tomato is red' as a premise in *reasoning* ('... therefore it is ripe'); it is also an achievement which is associated with a certain *qualitative experience* – an experience which feels a certain way, or has a certain **qualitative character**. This character, or quality, is its **quale**.

Philosophers use the existence of **qualia** to raise problems for theories of the mind in various ways. The following problems are usually raised as objections to **functionalism**; however, they can also be used to argue for **dualism** by attacking **materialism**. That's because these arguments talk about the possibility of a functional duplicate of you which either has different qualia or no qualia at all, showing that a functional organization is not *sufficient* (enough) for having conscious experience, and so doesn't explain conscious experience. However, on the assumption that any exactly physical duplicate of you is thereby also a functional duplicate, it follows that there could be an exact *physical* duplicate of you with different or absent qualia, showing that the total physical state of your body is not sufficient for having conscious experience and so doesn't explain conscious experience, or explain what it is to have a mind.

Qualia and Functionalism (I): absent qualia

The problem of **absent qualia** is the problem that there could be an exact **functional duplicate** of me that nevertheless did not experience qualia. This argument was introduced by Ned **Block**, who drew the further conclusion that, because the functional duplicate does not experience qualia, the functional duplicate does not have mental states at all; therefore, mental states cannot be merely functional states. This argument is also sometimes known as the **Chinese mind** argument, for reasons which will become clear:

> 'In this section I shall describe a class of devices that are prima facie embarrassments for all versions of functionalism in that they indicate functionalism is guilty of liberalism - classifying systems that lack mentality as having mentality. ...
>
> Imagine a body externally like a human body, say yours, but internally quite different. The neurons from sensory organs are connected to a bank of lights in a hollow cavity in the head. A set of buttons connects to the motor-output neurons. Inside the cavity resides a group of little men. Each has a very simple task: to implement a "square" of an adequate machine table that describes you. On one wall is a bulletin board on which is posted a state card, i.e., a card that bears a symbol designating one of the states specified in the machine table. Here is what the little men do: Suppose the posted card has a `G' on it. This alerts the little men who implement G squares - `G-men they call themselves. Suppose the light representing input I" goes on. One of the G-men has the following as his sole task: when the card reads `G' and the I" light goes on, he presses output button O,9, and changes the state card to 'M'. This G-man is called upon to exercise his task only rarely. In spite of the low level of intelligence required of each little man, the system as a whole manages to simulate you because the functional organization they have been trained to realize is yours. A Turing machine can be represented as a finite set of quadruples (or quintuples, if the output is divided into two parts): current state, current input; next state, next output. Each little man has the task corresponding to a single quadruple. Through the

efforts of the tittle men, the system realizes the same (reasonably adequate) machine table as you do and is thus functionally equivalent to you.

I shall describe a version of the homunculi-headed simulation, which has more chance of being nomologically possible. How many homunculi are required? Perhaps a billion are enough.

Suppose we convert the government of China to functionalism, and we convince its officials to realize a human mind for an hour. We provide each of the billion people in China (I chose China because it has a billion inhabitants) with a specially designed two-way radio that connects them in the appropriate way to other persons and to the artificial body mentioned in the previous example. We replace each of the little men with a citizen of China plus his radio. Instead of a bulletin board we arrange to have letters displayed on a series of satellites placed so that they can be seen from anywhere in China.

The system of a billion people communicating with one another plus satellites plays the role of an external "brain" connected to the artificial body by radio. There is nothing absurd about a person being connected to his brain by radio. Perhaps the day will come when our brains will be periodically removed for cleaning and repairs. Imagine that this is done initially by treating neurons attaching the brain to the body with a chemical that allows them to stretch like rubber bands, thereby assuring that no brain-body connections are disrupted. Soon clever businessmen discover that they can attract more customers by replacing the stretched neurons with radio links so that brains can be cleaned without inconveniencing the customer by immobilizing his body.

It is not at all obvious that the China-body system is physically impossible. It could be functionally equivalent to you for a short time, say an hour.

"But," you may object, "how could something be functionally equivalent to me for an hour? Doesn't my functional organization determine say, how I would react to doing nothing for a week but reading the Reader's Digest Remember that a machine table specifies a set of conditionals of the form: if the machine is in S; and receives input I, it emits output Ot and goes into S,. These conditionals are to be understood subjunctively. What gives a system a functional organization at a time is not just what it does at that time, but also the counterfactuals true of it at that time: what it would have done (and what its state transitions would have been) had it had a different input or been in a different state. If it is true of a system at time t that it would obey a given machine table no matter which of the states it is in and no matter which of the inputs it receives, then the system is described at t by the machine table (and realizes at t the abstract automaton specified by the table), even if it exists for only an instant. For the hour the Chinese system is "on,"it does have a set of inputs, outputs, and states of which such subjunctive conditionals are true. This is what makes any computer realize the abstract automaton that it realizes.

Of course, there are signals the system would respond to that you would not respond to e.g., massive radio interference or a flood of the Yangtze River. Such events might cause a malfunction, scotching the simulation, just as a bomb in a computer can make it fail to realize the machine table it was built to realize. But just as the computer without the bomb can realize the machine table, the system consisting

of the people and artificial body can realize the machine table so long as there are no catastrophic interferences, e.g., floods, etc.

"But;" someone may object, "there is a difference between a bomb in a computer and a bomb in the Chinese system, for in the case of the latter (unlike the former), inputs as specified in the machine table can be the cause of the malfunction. Unusual neural activity in the sense organs of residents of Chungking Province caused by a bomb or by a flood of the Yangtze can cause the system to go haywire."

Reply: The person who says what system he or she is talking about gets to say what signals count as inputs and outputs. I count as inputs and outputs only neural activity in the artificial body connected by radio to the people of China. Neural signals in the people of Chungking count no more as inputs to this system than input tape jammed by a saboteur between the relay contacts in the innards of a computer count as an input to the computer.

Of course, the object consisting of the people of China + the artificial body has other Turing-machine descriptions under which neural signals in the inhabitants of Chungking would count as inputs. Such a new system (i.e., the object under such a new Turing-machine description) would not be functionally equivalent to you. Likewise, any commercial computer can be redescribed in a way that allows tape jammed into its innards to count as inputs. In describing an object as a Turing machine, one draws a line between the inside and the outside. (If we count only neural impulses as inputs and outputs, we draw that line inside the body; if we count only peripheral stimulations as inputs, we draw that line at the skin.) In describing the Chinese system as a Turing machine, I have drawn the line in such a way that it satisfies a certain type of functional description-one that you also satisfy, and one that, according to functionalism, justifies attributions of mentality. Functionalism does not claim that every mental system has a machine table of a sort that justifies attributions of mentality with respect to every specification of inputs and outputs, but rather, only with respect to some specification.

Objection: The Chinese system would work too slowly. The kind of events and processes with which we normally have contact would pass by far too quickly for the system to detect them. Thus, we would be unable to converse with it, play bridge with it, etc.

Reply: It is hard to see why the system's time scale should matter. Is it really contradictory or nonsensical to suppose we could meet a race of intelligent beings with whom we could communicate only by devices such as time-lapse photography? When we observe these creatures, they seem almost inanimate. But when we view the time-lapse movies, we see them conversing with one another. Indeed, we find they are saying that the only way they can make any sense of us is by viewing movies greatly slowed down. To take time scale as all important seems crudely behavioristic.

What makes the homunculi-headed system (count the two systems as variants of a single system) just described a prima facie counterexample to (machine) functionalism is that there is prima facie doubt whether it has any mental states at all-especially whether it has what philosophers have variously called "qualitative states," "raw feels," or "immediate phenomenological qualities." (You ask: What is it that philosophers have called qualitative states? I answer, only half in jest: As Louis Armstrong said

> when asked what jazz is, "If you got to ask, you ain't never gonna get to know.") In Nagel's terms (1974), there is a prima facie doubt whether there is anything which it is like to be the homunculi-headed system... Call this argument the **absent qualia argument**.' Block, *Troubles with Functionalism*.

Qualia and Functionalism (II): the inverted spectrum

How do I know that the experience I have when I see a red object is qualitatively the same as the experience you have? Couldn't it be that *my* red-experience is actually the experience *you* have when you see green? It seems we could never tell for certain whether this was the case, since someone who was experiencing an **inverted spectrum** of this kind would still have the ability to name colours just as accurately as anyone else in the community. So it seems that there could be a **functional duplicate** of me who behaved in exactly the same way as I did, correctly identifying red things as red and so on, but whose qualia were completely different. Such a person, it seems, would be in completely *different* mental states, because his experiences would be different. But then sameness of functional organization does not entail sameness of mental state, and it cannot be claimed that mental states should be identified with functional states.

The possibility of an **inverted spectrum** was first raised by John **Locke** in 1689 (of course, Locke did not know about its relevance to functionalism, which would not be invented for nearly 300 years):

> Neither would it carry any Imputation of Falshood to our simple Ideas, if by the different Structure of our Organs, it were so ordered, That the same Object should produce in several Men's Minds different Ideas at the same time; v.g. if the Idea, that a Violet produced in one Man's Mind by his Eyes, were the same that a Marigold produces in another Man's, and vice versâ. For since this could never be known: because one Man's Mind could not pass into another Man's Body, to perceive, what Appearances were produced by those Organs; neither the Ideas hereby, nor the Names, would be at all confounded, or any Falshood be in either. For all Things, that had the Texture of a Violet, producing constantly the Idea, which he called Blue, and those which had the Texture of a Marigold, producing constantly the Idea, which he as constantly called Yellow, whatever those Appearances were in his Mind; he would be able as regularly to distinguish Things for his Use by those Appearances, and understand, and signify those distinctions, marked by the Names Blue and Yellow, as if the Appearances, or Ideas in his Mind, received from those two Flowers, were exactly the same, with the Ideas in other Men's Minds.'
>
> Locke, *Essay concerning Human Understanding*, II.xxxii.15

You should notice that the possibility of an inverted spectrum is *also* effective as an argument against **behaviourism**: the person whose colour-experiences are inverted in this way is behaviourally (as well as functionally) indistinguishable from someone with normal colour-vision.

One functionalist response to this argument is to claim that, as a matter of scientific fact, or **nomological necessity**, sameness of functional organization does entail sameness of qualia; there may be some as-yet-undiscovered scientific law which guarantees that any being in the same functional state as us will have the same kinds of experience. Thus inverted qualia could be described as **empirically impossible**. Some philosophers will object that this misses the point: even if (as it happens) the laws of nature rule out inverted spectra in *this* universe, all we need to show is that it is *logically* or *metaphysically* possible for some creature to be in the same functional state as us but with different qualia to establish that sameness of functional state does not guarantee sameness of functional state. (For a defence of this approach, see David Chalmers, *Absent*

Qualia, Fading Qualia, Dancing Qualia, online at http://cogprints.org/318/1/qualia.html).

An alternative functionalist response is to say that qualia, like other elements of our minds, should be individuated according to their functional role; so it is a *conceptual* truth that two people who are functionally indistinguishable will have the *same* qualia: if my experience of 'red' plays the same functional role that your experience of 'red' does, then our qualia are the same, regardless of the fact that in some way they 'feel' different. If qualia are identified with functional *roles*, then it might be no surprise that certain properties count as the *realizers* of those roles in me, and different properties count as realizers in other creatures whose makeup is different. Critics of this suggestion will respond that it misunderstands what qualia *are* – two people cannot have the same qualia *unless* things seem (subjectively) the same to them – so it makes no sense to offer a definition of qualia according to which two people could have the same qualia but different 'raw feels' to their experiences. (This response was introduced by Sidney **Shoemaker** – see his 1982 article *The Inverted Spectrum*, available online at
http://www.andrew.cmu.edu/user/kk3n/80-300/shoemaker-spectrum.pdf)

If you want to try an even bolder response to these anti-functionalist appeals to qualia, you could try the approach recommended by Dennett: 'there simply are no qualia at all.' Dennett's view is that someone who woke up one day experiencing the world as coloured in a completely different way would not be able to tell whether (i) their qualia were different from the ones they had had before or (ii) they were mis-remembering what their qualia used to be like. Since the concept of qualia is of no practical use in describing situations like these, Dennett concludes that the concept itself should be abandoned. (For more discussion, see Dennett, *Quining Qualia*, online at http://ase.tufts.edu/cogstud/papers/quinqual.htm – especially the section on the coffee tasters)

Week 11: Arguments for Property Dualism I: the Hard Problem of Consciousness

Previously we considered **substance dualism**. This was the view that the mind was a distinct non-physical kind of 'thing' – a 'mental substance' – which could not be explained, deduced, or 'reduced to' anything purely physical. Thus property dualism is opposed to **materialism / physicalism**, according to which there is nothing which is not part of the physical world. We noted that substance dualism can come in both **interactionist** and **epiphenomenalist** forms, depending on whether the mind itself is said to interact causally with the physical world (interactionism) or whether mental states and events are held to be caused by physical states of the brain but do not in turn cause anything themselves (epiphenomenalist substance dualism). Those two views are *similar* in so far as they both hold that the mind is ontologically distinct from the body or physical world; they are both motivated by the same kinds of arguments (e.g. Descartes' **indivisibility** and **conceivability** arguments), and they both claim that physical events in the brain can have nonphysical (mental) effects. However, they differ in that epiphenomenalists assert, and interactionists deny, that the physical world is a **causally closed system**; epiphenomenalists claim that mental states are **causally impotent** while interactionists deny this; and epiphenomenalists may even go so far as to say that mental states cannot even cause *other* mental states or events.

Now we are considering **property dualism**. In the first week of the course we considered the difference between mental and physical *properties*. (Think of a property as a 'way something is' or an 'attribute, feature or quality that something has'.) A paradigm physical property would be something like *being negatively charged* or *having mass*, while a paradigm mental property would be a property of your conscious experience such as *being painful*. **Qualia**, of course, are properties: they are **subjective or phenomenal features of mental states** (phenomenal just means 'to do with how it feels to us'), which are **introspectively accessible** to the person who has the mental state. So we can say that qualia are **the properties of what it is like to undergo the mental state in question**. A **property dualist** claims that mental properties are a distinct and fundamental *kind* of property: mental properties cannot be 'reduced to' or 'explained in terms of' physical properties, as the reductive materialist claims, and they cannot be explained in terms of a functional role, as the functionalist claims. So property dualism is opposed to both **materialism** and to **functionalism**.

While the property dualist accepts that there may be only one kind of 'stuff' or **substance** in the world (physical substance), so it is not correct to say that the mind is a numerically different entity from the brain, she will say that it is not true that everything there is, is physical, since there exist fundamental mental properties which cannot be explained away as somehow being physical properties 'in disguise': to use a frequently-quoted analogy, if God created this world, he would not automatically have created the mental properties simply by making the physical world and its properties; the mental properties would have to be added to the physical world as a separate act of creation. To weild a useful Latin tag: mental properties are **sui generis**, belonging to their own distinctive species or kind.

What influence do mental properties have on the physical world? Here the property dualist will chose between the same two approaches (interactionist and epiphenomenalist) that were available to the substance dualist. An **interactionist property dualist** will say that mental properties such as qualia can stand in causal relationships to physical states and events, and thus that mental properties can be **causally relevant** or **causally efficacious** with regard to what happens in the physical world. On the other hand, the

epiphenomenalist property dualist says that mental properties are caused by physical events, but do not go on to cause anything in turn: they are **causally impotent** or **causally irrelevant**. Thus the epiphenomenalist property dualist says that the felt properties of conscious experience are the result of what happens in the physical world, but do not influence what happens in the physical world. (A nice way of putting this asks us to imagine a steam train running along a track: the physical world would be the steam train and its track, while the mental properties are like the puffs of smoke emitted by the train engine: they are the result of the activity of the steam train but they do not make any difference to where it is going.)

Exercise: make revision charts explaining the key differences between (i) substance and property dualism; (ii) interactionist and epiphenomenalist dualism.

The 'Explanatory Gap' and the 'Hard Problem' of Consciousness

The phrase **explanatory gap** was coined by Joseph **Levine** in 1983 to talk about the way in which a materialist or scientific account of the mind would *fail* to be 'fully explanatory, with nothing crucial left out':

> 'Let me explain what I mean by an 'explanatory gap.' ...
>
> What is explained by learning that pain is the firing of C-fibers? Well, one might say that in fact quite a bit is explained. If we believe that part of the concept expressed by the term 'pain' is that of a state which plays a certain causal role in our interaction with the environment (e.g. it warns us of damage, it causes us to attempt to avoid situations we believe will result in it, etc.), [this] explains the mechanisms underlying the performance of these functions... However, there is more to our concept of pain than its causal role, there is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firing is *why pain should feel the way it does!*'
>
> Levine, *Materialism and Qualia: the Explanatory Gap*

The explanatory gap is probably best thought of as a challenge to **materialism**: any satisfactory materialist theory of mind must somehow explain everything that needs explaining about mental states; but the the existence of the explanatory gap highlights how difficult it will be to achieve the necessary explanation using only the resources of neuroscience. To use a over-used metaphor, how does the 'water' of physical matter give rise to the 'wine' of conscious experience? It seems no less miraculous than the events of the Wedding at Cana. (Further reading: Levine's original (short) paper is online at http://www.uoguelph.ca/~abailey/Resources/levine.pdf)

The phrase the **hard problem of consciousness** was coined by David Chalmers in the mid-1990s to describe the aspects of consciousness that it seems likely that current scientific methods in principle cannot explain. His aim in writing was partly to deflate the over-confidence of neuroscientists who believe that their continued success in solving various 'easy' problems of consciousness means that it is only a matter of time before they solve *all* the problems of consciousness. Read the following extract and answer the questions that follow:

> 'The word 'consciousness' is used in many different ways. It is sometimes used for the ability to discriminate stimuli, or to report information, or to monitor internal states, or to control behavior. We can think of these phenomena as posing the "easy problems" of consciousness. These are important phenomena, and there is much that is not understood about them, but the problems of explaining them have the character of puzzles rather than mysteries. There seems to be no deep

problem in principle with the idea that a physical system could be "conscious" in these senses, and there is no obvious obstacle to an eventual explanation of these phenomena in neurobiological or computational terms.

The hard problem of consciousness is the problem of experience. Humans beings have subjective experience: there is something it is like to be them. We can say that a being is conscious in this sense – or is phenomenally conscious, as it is sometimes put—when there is something it is like to be that being. A mental state is conscious when there is something it is like to be in that state. Conscious states include states of perceptual experience, bodily sensation, mental imagery, emotional experience, occurrent thought, and more. There is something it is like to see a vivid green, to feel a sharp pain, to visualize the Eiffel tower, to feel a deep regret, and to think that one is late. Each of these states has a phenomenal character, with phenomenal properties (or qualia) characterizing what it is like to be in the state.

There is no question that experience is closely associated with physical processes in systems such as brains. It seems that physical processes give rise to experience, at least in the sense that producing a physical system (such as a brain) with the right physical properties inevitably yields corresponding states of experience. But how and why do physical processes give rise to experience? Why do not these processes take place "in the dark," without any accompanying states of experience? This is the central mystery of consciousness.

What makes the easy problems easy? For these problems, the task is to explain certain behavioral or cognitive functions: that is, to explain how some causal role is played in the cognitive system, ultimately in the production of behavior. To explain the performance of such a function, one need only specify a mechanism that plays the relevant role. And there is good reason to believe that neural or computational mechanisms can play those roles.

What makes the hard problem hard? Here, the task is not to explain behavioral and cognitive functions: even once one has an explanation of all the relevant functions in the vicinity of consciousness – discrimination, integration, access, report, control—there may still remain a further question: why is the performance of these functions accompanied by experience? Because of this, the hard problem seems to be a different sort of problem, requiring a different sort of solution.

A solution to the hard problem would involve an account of the relation between physical processes and consciousness, explaining on the basis of natural principles how and why it is that physical processes are associated with states of experience. A reductive explanation of consciousness will explain this wholly on the basis of physical principles that do not themselves make any appeal to consciousness. A materialist (or physicalist) solution will be a solution on which consciousness is itself seen as a physical process. A nonmaterialist (or nonphysicalist) solution will be a solution on which consciousness is seen as nonphysical (even if closely associated with physical processes). A nonreductive solution will be one on which consciousness (or principles involving consciousness) is admitted as a basic part of the explanation.

It is natural to hope that there will be a materialist solution to the hard problem and a reductive

explanation of consciousness, just as there have been reductive explanations of many other phenomena in many other domains. But consciousness seems to resist materialist explanation in a way that other phenomena do not.'

Chalmers, *Consciousness and its place in nature* (2003)

Questions:

1) In a word, what is the 'hard problem of consciousness'?
2) What is the 'central mystery of consciousness'?
3) Why are the easy problems easy?
4) Why is the hard problem hard?
5) What would a solution to the hard problem involve?

Commentary

Chalmers bases his idea of the 'hard problem of consciousness' on what he thinks is a defining feature of scientific / physical explanation: that it explains how functions are performed in terms of 'mechanisms' which explain *how* the process leads to the end result. Here's how he puts that point in his original paper, *Facing up to the Problem of Consciousness* (1995):

> 'Throughout the higher-level sciences, reductive explanation works in just this way. To explain the gene, for instance, we needed to specify the mechanism that stores and transmits hereditary information from one generation to the next. It turns out that DNA performs this function; once we explain how the function is performed, we have explained the gene. To explain life, we ultimately need to explain how a system can reproduce, adapt to its environment, metabolize, and so on. All of these are questions about the performance of functions, and so are well-suited to reductive explanation.'

Chalmers *also* claims that the problem of consciousness is not that kind of question: explaining *why* a certain process is accompanied by conscious experience is not the same as explaining *how* that process is carried out. He summarizes his argument as follows, then adds an additional premise to get us from the claim that consciousness cannot be explained in physical terms to the conclusion that consciousness 'is not itself physical' and thus that physicalism is false:

> 'We can call this the explanatory argument:
>
> (1) Physical accounts explain at most structure and function.
>
> (2) Explaining structure and function does not suffice to explain consciousness; so
>
> ————————-
>
> (3) No physical account can explain consciousness.
>
> If this is right, then while physical accounts can solve the easy problems (which involve only explaining functions), something more is needed to solve the hard problem. It would seem that no reductive explanation of consciousness could succeed. And if we add the premise that what cannot be physically explained is not itself physical (this can be considered an additional final step of the explanatory argument), then materialism about consciousness is false, and the natural world contains more than the physical world.' Chalmers, *Consciousness and...*

However, a major criticism of the Hard Problem is that the mere fact that something *seems* inexplicable in physical terms does not mean that it *is* inexplicable in physical terms. For example, in the 18th century some biologists and chemists argued that the difference between living things (plants and animals) and others could only be explained by a special 'vital force' which enabled living things to grow and multiply. Here's how Chalmers responds to that objection:

> '[Critics of the Hard Problem] often press a different sort of analogy, holding that at various points in the past, thinkers held that there was an analogous epistemic gap for other phenomena, but that these turned out to be physically explained. For example, Dennett (1996) suggests that a vitalist might have held that there was a further "hard problem" of life over and above explaining the biological function, but that this would have been misguided.

On examining the cases, however, the analogies do not support the type-A materialist. Vitalists typically accepted, implicitly or explicitly, that the biological functions in question were what needed explaining. Their vitalism arose because they thought that the functions (adaptation, growth, reproduction, and so on) would not be physically explained. So this is quite different from the case of consciousness. The disanalogy is very clear in the case of [C.D.] Broad. Broad was a vitalist about life, holding that the functions would require a non-mechanical explanation. But at the same time, he held that in the case of life, unlike the case of consciousness, the only evidence we have for the phenomenon is behavioral, and that "being alive" means exhibiting certain sorts of behavior. Other vitalists were less explicit, but very few of them held that something more than the functions needed explaining (except consciousness itself, in some cases). If a vitalist had held this, the obvious reply would have been that there is no reason to believe in such an explanandum. So there is no analogy here.

So these arguments by analogy have no force for the type-A materialist. In other cases, it was always clear that structure and function exhausted the apparent explananda, apart from those tied directly to consciousness itself. So the type-A materialist needs to address the apparent further explanandum in the case of consciousness head on: either flatly denying it, or giving substantial arguments to dissolve it.' Chalmers, *Consciousness and...*

Further viewing:

Daniel Dennett deflates consciousness at http://www.youtube.com/watch?v=vkaS5JWZ1hY.

Ned Block on conscious as an illusion at http://www.youtube.com/watch?v=N6SbPPL8tOI.

Week 11: Arguments for Property Dualism II: Zombies!

Zombies

Viewing: Chalmers on Zombies at (http://www.youtube.com/watch?v=NK1Yo6VbRoo)

A philosophical **zombie** is physically, functionally, and behaviourally indistinguishable from a normal human being: it responds in exactly the same way as a normal person to external stimuli, showing all the outer signs of pain, emotion, etc. Moreover, it has a normally-functioning human brain. The only thing different about the zombie is that it has *no conscious experience whatsoever*. It might claim to be in great pain, and show all the outward signs of pain, but it is in fact experiencing nothing – there is nothing 'it is like' to be the zombie, as the zombie has no experiences.

One use of philosophical zombies is to raise a version of the **problem of other minds**: how do I know that other people really have conscious experience, rather than merely claiming to do so? But zombies were originally introduced to philosophy to serve a different purpose: to argue that 'no explanation given wholly in physical terms can ever account for the emergence of conscious experience' (Chalmers). In other words, Zombies form part of an argument against physicalism, and in favour of some form of dualism: it is not the case that the physical facts fix all the facts there are, since there are some facts (facts about consciousness) which are not dependent on physical facts in any way. If a **zombie world** (a world exactly the same as this one, except that everyone in it is a zombie) is metaphysically possible, then it is metaphysically possible to have all the physical facts without thereby getting the facts about consciousness. Here's how Chalmers explains the argument:

> '… it is conceivable that there be a system that is physically identical to a conscious being, but that lacks at least some of that being's conscious states. Such a system might be a zombie: a system that is physically identical to a conscious being but that lacks consciousness entirely. It might also be an invert, with some of the original being's experiences replaced by different experiences, or a partial zombie, with some experiences absent, or a combination thereof. These systems will look identical to a normal conscious being from the third-person perspective: in particular, their brain processes will be molecule-for-molecule identical with the original, and their behavior will be indistinguishable. But things will be different from the first-person point of view. What it is like to be an invert or a partial zombie will differ from what it is like to be the original being. And there is nothing it is like to be a zombie.
>
> There is little reason to believe that zombies exist in the actual world. But many hold that they are at least conceivable: we can coherently imagine zombies, and there is no contradiction in the idea that reveals itself even on reflection. As an extension of the idea, many hold that the same goes for a zombie world: a universe physically identical to ours, but in which there is no consciousness. Something similar applies to inverts and other duplicates.
>
> From the conceivability of zombies, proponents of the argument infer their metaphysical possibility. Zombies are probably not naturally possible: they probably cannot exist in our world, with its laws of nature. But the argument holds that zombies could have existed, perhaps in a very different sort of universe. For example, it is sometimes suggested that God could have created a zombie world, if he had so chosen. From here, it is inferred that consciousness must be nonphysical. If there is a

metaphysically possible universe that is physically identical to ours but that lacks consciousness, then consciousness must be a further, nonphysical component of our universe. If God could have created a zombie world, then (as Kripke puts it) after creating the physical processes in our world, he had to do more work to ensure that it contained consciousness.

We can put the argument, in its simplest form, as follows:

(1) It is conceivable that there be zombies

(2) If it is conceivable that there be zombies, it is metaphysically possible that there be zombies.

(3) If it is metaphysically possible that there be zombies, then consciousness is non-physical.

————————-

(4) Consciousness is nonphysical.'

Chalmers, *Consciousness and...*

Commentary and criticism

A first criticism of Chalmer's argument (which he is fully aware of) is that **a 'zombie' world is not conceivable**: in other words, that although we *think* we are coherently imagining a situation in which creatures physically just like us exist, but lack conscious experience, we are actually *not* really conceiving of that at all. That's the view taken by Daniel Dennett, who suggests that the attempt to conceive of a functioning human being without consciousness is based on a misunderstanding of what consciousness *is*:

'If consciousness is (as I argue) not a single wonderful separable thing ('experiential sensitivity') but a huge complex of many different informational capacities that individually arise for a wide variety of reasons, there is no reason to suppose that 'it' is something that stands in need of its own separable status as fitness-enhancing. It is not a separate organ or a separate medium or a separate talent.

To see the fallacy, consider the parallel question about what the adaptive advantage of health is. Consider 'health inessentialism': for any bodily activity *b*, performed in any domain *d*, even if we need to be healthy to engage in it (e.g. pole vaulting, swimming the English Channel, climbing Mount Everest), it could in principle be engaged in by something that wasn't healthy at all. So what is health for? Such a mystery! But the mystery would arise only for someone who made the mistake of supposing that health was some additional thing that could be added or subtracted to the proper workings of all the parts. In the case of health we are not apt to make such a simple mistake, but there is a tradition of supposing just this in the case of consciousness. Supposing that by an act of stipulative imagination you can remove consciousness while leaving all cognitive systems intact — a quite standard but entirely bogus feat of imagination — is like supposing that by an act of stipulative imagination, you can remove health while leaving all bodily functions and powers intact. If you think you can imagine this, it's only because you are confusedly imagining some health-module that might or might not be present in a body. Health isn't that sort of thing, and neither is consciousness.'

Dennett, *The Unimagined Preposterousness of Zombies*

A response to this on behalf of Chalmers might take the form of saying that Dennett *begs the question* against the Zombie argument (presupposes what he is trying to establish), by claiming that 'consciousness is … a huge complex of many different informational capacities'. If that is right, then obviously it makes no sense to ask whether some creature could have all those capacities and yet fail to be conscious, so it is *true* that zombies cannot be conceived. But Chalmers rejects that account of consciousness in terms of 'capacities': according to him, to be conscious a creature must also have subjective, phenomenal experience: there must be 'something it is like' to be that creature; that creature must be aware of **qualia**. The 'zombie' scenario is supposed to be one in which a creature is functionally equivalent to us but lacks qualia; Dennett suggests that any creature functionally equivalent to us would count as having 'consciousness', but that just ignores what Chalmers takes to be a necessary condition for consciousness, namely having qualia. Without an independent *argument* to the effect that conscious requires only 'capacities' and not qualia, Dennett's objection will not convince us.

A second criticism is that zombies may be conceivable, but they are not, after all, genuinely possible: what we can 'conceive of' tells us about the limits of our minds, not about what is genuinely possible. for example, it is *conceivable* that Pythagoras' theorem is false – but it isn't logically possible. You have already met Chalmer's response to this criticism when studying Descartes' **conceivability** argument, but here it is again:

> 'The question is not whether it is plausible that zombies could exist in our world... the question is whether the notion of a zombie is conceptually coherent. The mere intelligibility of the notion is enough to establish the conclusion.
>
> Arguing for a logical possibility is not entirely straightforward. How, for example, would one argue that a mile-high unicycle is logically possible? It just seems obvious. Although no such thing exists in the real world, the description certainly appears to be coherent... I confess that the logical possibility of zombies seems equally obvious to me. A zombie is just something physically identical to me, but which has no conscious experience – all is dark inside. While this is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description... If no reasonable analysis of the terms in question points towards a contradiction, or even makes the existence of a contradiction possible, then there is a natural assumption in favor of logical possibility.' Chalmers, *The Conscious Mind*

Chalmers can also draw some support here by making a distinction between **nomological possibility** on the one hand, and **logical** or **metaphysical possibility** on the other. The laws of nature determine what is nomologically possible ('nomos' means 'law' in Greek): so, for example, it is nomologically impossible for anything to travel faster than 299,792,458 metres per second, as that's the speed of light in this, the actual world. But it's certainly *possible* in some sense that things could go faster than that, because the laws of physics could have been different (they are **contingent**), and so any fact that depends on them 'could have been otherwise'. Chalmers is not arguing that, because zombies are conceivable, they are nomologically possible (that's like arguing that, because it's conceivable that something travels faster than light, something *can* travel faster than light). All he needs to do is show that the conceivability of zombies entails their possibility in the wider 'logical' or 'metaphysical' sense, and it's hard to see how something could be *impossible* in that broader sense unless there was some hidden logical contradiction in the way the possibility was described.

A third, final objection is familiar from the discussion of Descartes' conceivability argument: that **what is logically possible tells us nothing about reality**. As before, we can introduce a distinction between **logical** and **metaphysical possibility** to make sense of this. If something *really could have been the case* in some possible world, then it is 'metaphysically possible', whereas if something is consistent with the laws of logic, it is 'logically possible'. Chalmers in fact changed his mind about how to present the kind of possibility involved in his 'zombie' argument: originally he had claimed he had established the **logical possibility of a physical duplicate of this world but without consciousness / qualia**, but in the more recent presentation quoted above he claims that the 'zombie' world is **metaphysically possible** – it really could have been the case.

Clearly metaphysical possibility is the correct kind of possibility for the argument to work, but it is not clear that you can derive a metaphysical possibility from a logical possibility. For example, in the work of Saul **Kripke** (who you read in connection with type-identity), it is argued that there are **necessities** in this world which are not knowable *a priori* (just through thinking) and so are **metaphysically necessary** despite not being **logically necessary** (because if you could derive them simply from the laws of logic, you'd be able to do that just by thinking). To turn that on its head: showing that something is **logically possible** doesn't show that it's **metaphysically possible** – because there might be some other, non-*a priori*, non-logical necessity which stops it from being the case. To take just one of Kripke's examples: it's metaphysically impossible that *this very table* could have been made out of ice, since *this* table (the one I'm talking about) could only be this table if it were made out of the same material it actually is made out of. But it is consistent with the laws of logic that this table could have been made out of anything at all: logic just doesn't lay down rules about what tables can be made out of!

Further reading:
Chalmers, *Consciousness and its Place in Nature*
'Zombies on the Web' at http://consc.net/zombies.html.
Stanford Encyclopedia of Philosophy, 'Zombies' (http://plato.stanford.edu/entries/zombies/)

Week 13: Arguments for Property Dualism III: the Knowledge Argument and 'Mary'

The **knowledge argument** is elegantly explained in Frank Jackson's 1982 article *Epiphenomenal Qualia:*

> 'It is undeniable that the physical, chemical and biological sciences have provided a great deal of information about the world we live in and about ourselves. I will use the label 'physical information' for this kind of information, and also for information that automatically comes along with it. For example, if a medical scientist tells me enough about the processes that go on in my nervous system, and about how they relate to happenings in the world around me, to what has happened in the past and is likely to happen in the future, to what happens to other similar and dissimilar organisms, and the like, he or she tells me — if I am clever enough to fit it together appropriately — about what is often called the functional role of those states in me (and in organisms in general in similar cases). This information, and its kin, I also label 'physical'.
>
> I do not mean these sketchy remarks to constitute a definition of 'physical information', and of the correlative notions of physical property, process, and so on, but to indicate what I have in mind here. It is well known that there are problems with giving a precise definition of these notions, and so of the thesis of Physicalism that all (correct) information is physical information. But — unlike some — I take the question of definition to cut across the central problems I want to discuss in this paper.
>
> I am what is sometimes known as a "qualia freak." I think that there are certain features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes. Tell me everything physical there is to tell about what is going on in a living brain, the kind of states, their functional role, their relation to what goes on at other times and in other brains, and so on and so forth, and be I as clever as can be in fitting it all together, you won't have told me about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky.
>
> There are many qualia freaks, and some of them say that their rejection of Physicalism is an unargued intuition. I think that they are being unfair to themselves. They have the following argument. Nothing you could tell of a physical sort captures the smell of a rose, for instance. Therefore, Physicalism is false. By our lights this is a perfectly good argument. It is obviously not to the point to question its validity, and the premise is intuitively obviously true both to them and to me.
>
> I must, however, admit that it is weak from a polemical point of view. There are, unfortunately for us, many who do not find the premise intuitively obvious. The task then is to present an argument whose premises are obvious to all, or at least to as many as possible. This I try to do in section I with what I will call "the Knowledge argument." ...
>
> People vary considerably in their ability to discriminate colors. Suppose that in an experiment to catalog this variation Fred is discovered. Fred has better color vision than anyone else on record; he makes every discrimination that anyone has ever made, and moreover he makes one that we cannot even begin to make. Show him a batch of ripe tomatoes and he sorts them into two roughly equal groups and does so with complete consistency. That is, if you blindfold him, shuffle the tomatoes up, and then remove the blindfold and ask him to sort them out again, he sorts them into exactly the same two groups.

We ask Fred how he does it. He explains that all ripe tomatoes do not look the same color to him, and in fact that this is true of a great many objects that we classify together as red. He sees two colors where we see one, and he has in consequence developed for his own use two words'red1' and 'red2' to mark the difference. Perhaps he tells us that he has often tried to teach the difference between red1 and red2 to his friends but has got nowhere and has concluded that the rest of the world is red1-red2 color-blind —or perhaps he has had partial success with his children; it doesn't matter. In any case he explains to us that it would be quite wrong to think that because 'red' appears in both 'red1'and 'red2' that the two colors are shades of the one color. He only uses the common term 'red' to fit more easily into our restricted usage. To him red1 and red2 are as different from each other and all the other colors as yellow is from blue. And his discriminatory behavior bears this out: he sorts red1 from red2 tomatoes with the greatest of ease in a wide variety of viewing circumstances. Moreover, an investigation of the physiological basis of Fred's exceptional ability reveals that Fred's optical system is able to separate out two groups of wavelengths in the red spectrum as sharply as we are able to sort out yellow from blue.

I think that we should admit that Fred can see, really see, at least one more color than we can; red1 is a different color from red2. We are to Fred as a totally red-green color-blind person is to us. H. G. Wells' story "The country of the blind" is about a sighted person in a totally blind community. This person never manages to convince them that he can see, that he has an extra sense. They ridicule this sense as quite inconceivable, and treat his capacity to avoid falling into ditches, to win fights and so on as precisely that capacity and nothing more. We would be making their mistake if we refused to allow that Fred can see one more color than we can.

What kind of experience does Fred have when he sees red1 and red2? What is the new color or colors like? We would dearly like to know but do not; and it seems that no amount of physical information about Fred's brain and optical system tells us. We find out perhaps that Fred's cones respond differentially to certain light waves in the red section of the spectrum that make no difference to ours (or perhaps he has an extra cone) and that this leads in Fred to a wider range of those brain states responsible for visual discriminatory behavior. But none of this tells us what we really want to know about his color experience. There is something about it we don't know. But we know, we may suppose, everything about Fred's body, his behavior and dispositions to behavior and about his internal physiology, and everything about his history and relation to others that can be given in physical accounts of persons. We have all the physical information. Therefore, knowing all this is not knowing everything about Fred. It follows that Physicalism leaves something out.

To reinforce this conclusion, imagine that as a result of our investigations into the internal workings of Fred we find out how to make everyone's physiology like Fred's in the relevant respects; or perhaps Fred donates his body to science and on his death we are able to transplant his optical system into someone else — again the fine detail doesn't matter. The important point is that such a happening would create enormous interest. People would say "At last we will know what it is like to see the extra color, at last we will know how Fred has differed from us in the way he has struggled to tell us about for so long." Then it cannot be that we knew all along all about Fred. But ex hypothesi we did know all

along everything about Fred that features in the physicalist scheme; hence the physicalist scheme leaves something out.

Put it this way. After the operation, we will know more about Fred and especially about his color experiences. But beforehand we had all the physical information we could desire about his body and brain, and indeed everything that has ever featured in physicalist accounts of mind and consciousness. Hence there is more to know than all that. Hence Physicalism is incomplete.

Fred and the new color(s) are of course essentially rhetorical devices. The same point can be made with normal people and familiar colors. Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red','blue', and so on. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'. (It can hardly be denied that it is in principle possible to obtain all this physical information from black and white television, otherwise the Open University would of necessity need to use color television.)

What will happen when Mary is released from her black and white room or is given a color television monitor? Will she learn anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had all the physical information. Ergo there is more to have than that, and Physicalism is false.

Clearly the same style of Knowledge argument could be deployed for taste, hearing, the bodily sensations and generally speaking for the various mental states which are said to have (as it is variously put) raw feels, phenomenal features or qualia. The conclusion in each case is that the qualia are left out of the physicalist story. And the polemical strength of the Knowledge argument is that it is so hard to deny the central claim that one can have all the physical information without having all the information there is to have.'

Jackson's argument is simple but effective: you could know everything there is to know about the physical world, and still not know about qualia; so qualia are not part of the physical world – even though they *do* exist. So not everything is physical. In fact, Jackson is more specific: it's not that there is something it's like to *be* Fred, or something it's like for Mary to see colour, which is inaccessible to us: instead, there is in each case 'something about their experience' – namely a property of what their experience is like – which cannot be deduced from the purely physical information we have:

'When I complained that all the physical knowledge about Fred was not enough to tell us what his special colour experience was like, I was not complaining that we weren't finding out what it is like to be Fred. *I was complaining that there is something about his experience, a property of it, of which we were left ignorant*. And if and when we come to know what this property is we still will not know what

it is like to be Fred, but we will know more about him. No amount of knowledge about Fred, be it physical or not, amounts to knowledge "from the inside" concerning Fred. We are not Fred. There is thus a whole set of items of knowledge expressed by forms of words like 'that it is I myself who is . . .' which Fred has and we simply cannot have because we are not him.'

So Jackson can run his 'Mary' argument *both* as a general argument against physicalism, *and* as an argument for **property dualism**. In the first form, we argue thus:

(1) Mary knows all the physical facts (before she leaves the room)

(2) Mary does not know all the facts (because she does not know facts about the subjective character or qualia associated with colour-experiences).

(3) Therefore, the physical facts are not all the facts.

As an argument for property dualism, the argument runs like this:

(1) Mary knows about all the physical properties (before she leaves the room)

(2) Mary does not know about all the properties (because she does not know about the subjective character or qualia associated with colour-experiences).

(3) Therefore, the physical properties are not all the properties.

Epiphenomenalism again

Jackson originally defended **epiphenomenalist property dualism**: his view was that, although some mental states might be 'causally efficacious', there was no good reason to believe that **qualia** have any role in physical causation. He considers, and rejects, three arguments which attempt to show that qualia do have a causal role. Read the following extract, and make your own chart summarizing the arguments and Jackson's responses to them:

'Three reasons are standardly given for holding that a quale like the hurtfulness of a pain must be causally efficacious in the physical world, and so, for instance, that its instantiation must sometimes make a difference to what happens in the brain. None, I will argue, has any real force. (I am much indebted to Alee Ryslop and John Lucas for convincing me of this.)

(i) It is supposed to be just obvious that the hurtfulness of pain is partly responsible for the subject seeking to avoid pain, saying 'It hurts' and so on. But, to reverse Hume, anything can fail to cause anything. No matter how often B follows A, and no matter how initially obvious the causality of the connection seems, the hypothesis that A causes B can be overturned by an over-arching theory which shows the two as distinct effects of a common underlying causal process. To the untutored the image on the screen of Lee Marvin's fist moving from left to right immediately followed by the image of John Wayne's head moving in the same general direction looks as causal as anything. And of course throughout countless Westerns images similar to the first are followed by images similar to the second. All this counts for precisely nothing when we know the over-arching theory concerning how the relevant images are both effects of an underlying causal process involving the projector and the film. The epiphenomenalist can say exactly the same about the connection between, for example, hurtfulness and behaviour. It is simply a consequence of the fact that certain happenings in the brain cause both.

(ii) The second objection relates to Darwin's Theory of Evolution. According to natural selection the traits that evolve over time are those conducive to physical survival. We may assume that qualia evolved over time - we have them, the earliest forms of life do not - and so we should expect qualia to be conducive to survival. The objection is that they could hardly help us to survive if they do nothing to the physical world.

The appeal of this argument is undeniable, but there is a good reply to it. Polar bears have particularly thick, warm coats. The Theory of Evolution explains this (we suppose) by pointing out that having a thick, warm coat is conducive to survival in the Arctic. But having a thick coat goes along with having a heavy coat, and having a heavy coat is not conducive to survival. It slows the animal down. Does this mean that we have refuted Darwin because we have found an evolved trait - having a heavy coat – which is not conducive to survival? Clearly not. Having a heavy coat is an unavoidable concomitant of having a warm coat (in the context, modern insulation was not available), and the advantages for survival of having a warm coat outweighed the disadvantages of having a heavy one.

The point is that all we can extract from Darwin's theory is that we should expect any evolved characteristic to be either conducive to survival or a by-product of one that is so conducive. The epiphenomenalist holds that qualia fall into the latter category. They are a by-product of certain brain processes that are highly conducive to survival.

(iii) The third objection is based on a point about how we come to know about other minds. We know about other minds by knowing about other behaviour, at least in part. The nature of the inference is a matter of some controversy, but it is not a matter of controversy that it proceeds from behaviour. That is why we think that stones do not feel and dogs do feel. But, runs the objection, how can a person's behaviour provide any reason for believing he has qualia like mine, or indeed any qualia at all, unless this behaviour can be regarded as the outcome of the qualia. Man Friday's footprint was evidence of Man Friday because footprints are causal outcomes of feet attached to people. And an epiphenomenalist cannot regard behaviour, or indeed anything physical, as an outcome of qualia.

But consider my reading in The Times that Spurs won. This provides excellent evidence that The Telegraph has also reported that Spurs won, despite the fact that (I trust) The Telegraph does not get the results from The Times. They each send their own reporters to the game. The Telegraph's report is in no sense an outcome of The Times', but the latter provides good evidence for the former nevertheless. The reasoning involved can be reconstructed thus. I read in The Times that Spurs won. This gives me reason to think that Spurs won because I know that Spurs' winning is the most likely candidate to be what caused the report in The Times. But I also know that Spurs' winning would have had many effects, including almost certainly a report in The Telegraph. I am arguing from one effect back to its cause and out again to another effect. The fact that neither effect causes the other is irrelevant. Now the epiphenomenalist allows that qualia are effects of what goes on in the brain. Qualia cause nothing physical but are caused by something physical. Hence the epiphenomenalist can argue from the behaviour of others to the qualia of others by arguing from the behaviour of others back to its causes in the brains of others and out again to their qualia.'

Finally, Jackson raises a question about the physicalist's insistence that everything *must* be able to be explained in physical terms: why should we be so 'optimistic' as to expect that this will be the case? Could it not be that the ways of thinking associated with physical science are just an expression of our own limited perspective on reality? To make that point vivid, he introduces the idea of the 'slugists':

> 'suppose we discovered living on the bottom of the deepest oceans a sort of sea slug which manifested intelligence. Perhaps survival in the conditions required rational powers. Despite their intelligence, these sea slugs have only a very restricted conception of the world by comparison with ours, the explanation for this being the nature of their immediate environment. Nevertheless they have developed sciences which work surprisingly well in these restricted terms. They also have philosophers, called slugists. Some call themselves tough-minded slugists, others confess to being soft-minded slugists.
>
> The tough-minded slugists hold that the restricted terms (or ones pretty like them which may be introduced as their sciences progress) suffice in principle to describe everything without remainder. These tough-minded slugists admit in moments of weakness to a feeling that their theory leaves something out. They resist this feeling and their opponents, the soft-minded slugists, by pointing out – absolutely correctly – that no slugist has ever succeeded in spelling out how this mysterious residue fits into the highly successful view that their sciences have and are developing of how their world works.
>
> Our sea slugs don't exist, but they might. And there might also exist super beings which stand to us as we stand to the sea slugs. We cannot adopt the perspective of these super beings, because we are not them, but the possibility of such a perspective is, I think, an antidote to excessive optimism.'

Week 14: Responses to the Knowledge Argument

A first response to the knowledge argument is to say that **Mary gains no new propositional knowledge** (knowledge *that* something is the case); instead she gains **acquaintance knowledge** (like being introduced to a person for the first time), or **ability knowledge** (i.e. she gains a particular kind of ability to do something, not extra knowledge of facts about the world).

One way to do this is to say that all Mary gains on leaving the room is an introduction to properties of conscious experience which she has not encountered before, and this is not really propositional knowledge. But Jackson will respond that she *does* learn new facts: facts about the mental life of *other people*:

> 'after Mary sees her first ripe tomato, she will realize how impoverished her conception of the mental life of others has been all along. She will realize that there was, all the time she was carrying out her laborious investigations into the neurophysiologies of others and into the functional roles of their internal states, something about these people she was quite unaware of. All along their experiences (or many of them, those got from tomatoes, the sky, . . .) had a feature conspicuous to them but until now hidden from her (in fact, not in logic).'

Alternatively, it might be claimed that what Mary gains is a certain kind of *ability*: the ability to represent the colour red to herself, or the ability to recognize red things just by looking at them. Here's how Jackson replies to that version of the objection:

> Now it is certainly true that Mary will acquire abilities of various kinds after her release. She will, for instance, be able to imagine what seeing red is like, be able to remember what it is like, and be able to understand why her friends regarded her as so deprived (something which, until her release, had always mystified her). But is it plausible that that is all she will aquire? Suppose she received a lecture on skepticism about other minds while she was incarcerated. On her release she sees a ripe tomato in normal conditions, and so has a sensation of red. Her first reaction is to say that she now knows more about the kind of experiences others have when looking at ripe tomatoes. She then remembers the lecture and starts to worry. Does she really know more about what their experiences are like, or is she indulging in a wild generalization from one case? In the end she decides she does know, and that skepticism is mistaken (even if, like so many of us, she is not sure how to demonstrate its errors). What was she to-ing and fro-ing about-her abilities? Surely not; her representational abilities were a known constant throughout. What else then was she agonizing about than whether or not she had gained factual knowledge of others? There would be nothing to agonize about if ability was all she acquired on her release.'

A **second** objection to the 'Mary' argument is to deny the first premise: it is not true that Mary, inside the black-and-white room, knows all the physical facts, because **all physical knowledge would include knowledge of qualia**. What the Knowledge Argument tells us, then, is simply that you can't acquire all the physical facts sitting inside a black-and-white room; you have to go outside and see some red things as well. But this seems implausible. If facts about qualia really are physical facts, then it should be possible to state these facts in the language of physics, neuroscience, chemistry, biology or something else like that. And if these facts *are* stateable in those terms then there is no reason why they couldn't be learnt *inside* the black-and-white room. In other words, if facts about qualia *were* physical facts, then Mary should be able to learn them by listening to

a lecture or watching a science documentary on her black-and-white television. But she can't; so they're not physical facts.

Here's an alternative way to run the objection: maybe facts about qualia are *consequences* of physical facts: Mary does in fact know all the basic physical facts, but is unaware of some of the consequences of those facts. So physicalism can still be true: everything is *fundamentally* physical, but some physical facts have consequences which will not be obvious to everyone, since people might lack the ability to make the relevant inferences. To answer this, later formulations of the Knowledge Argument add the premise that Mary is a 'perfect reasoner' who can deduce all the consequences of the physical facts she learns inside the black-and-white room. Suppose that's true: Mary, in her room, knows all the physical facts *and* all the consequences of the physical facts. The supporter of the argument will say that *even in this case* Mary wouldn't know facts about qualia, since no amount of reasoning about the physical nature of our visual system will tell us that the experience of seeing red things is accompanied by *this* quale rather than any other.

A **third** objection claims that **there is more than one way of knowing the same physical fact**. This is also known as the 'new knowledge / old fact' view. Here the argument relies on pointing out that you can think of the same thing under two different concepts: for example the colour 'red' can be conceptualized both in physical terms (in terms of wavelengths of light and so on) *and* in 'phenomenal' terms (in terms of how it appears to us). So we can conceptualize one and the same fact in two different ways, depending on whether we think of it using our physical or the phenomenal concepts. So Mary does not gain knowledge of any new facts; she just acquires a new way to conceptualize the facts about colour that she already knows. This objection is very popular among philosophers, partly because it explains why it *seems* to Mary that she is gaining new knowledge – a new conceptualization of an old fact will seem like new knowledge.

However, it is not clear that it can succeed. Qualia can *only* be conceptualized in 'phenomenal' terms, because qualia are, by definition, phenomenal features of mental states. So if Mary gains knowledge of a fact about qualia, like '*this* is the quale associated with experiences of red things', it is hard to see how someone could argue that that fact is just a re-conceptualization of a physical fact: if it were, there would have to be some way of conceptualizing qualia themselves in purely physical terms, and there isn't.

Finally, some philosophers respond to the Knowledge Argument by claiming that **qualia do not exist, and so Mary gains no propositional knowledge**. This is a view associated with Daniel **Dennett**. One way of motivating it is to think about what qualia are supposed to be: they are thought of as properties of experiences (or other mental states) which are **intrinsic** (they belong to the experience itself), **non-representational** (they do not directly represent something in the world), **directly accessible** by the subject and **subjective** in the sense that, if it seems to someone that their experience has a particular quale, it *does* have that quale. One way to raise a question about the existence of qualia is to ask whether *experiences* themselves have properties, or whether we experience *things* in the world as having properties. Here's how Tim **Crane** explains that point:

> 'It is not at all obvious that when we learn what it is like to taste retsina, we are learning about a property of an experience. Isn't it slightly more obvious, at least at first sight, that we are learning something about retsina: viz., what it tastes like, or what it is like to taste it? Yet many philosophers do take such knowledge to be obvious.' Crane, *The Origins of Qualia*

Dennett attacks qualia on the grounds that there is no such thing as 'the way something seems to you' independently of how you react to that experience. Some of his examples bear repeating:

> 'Imagine now … that someone offers me a pill to cure my loathing for cauliflower . He promises that after I swallow this pill cauliflower will taste exactly the same to me as it always has, but I will like that taste. "Hang on," I might reply, "I think you may have just contradicted yourself." But in any event I take the pill and it works . I become an instant cauliflower - appreciator … Of course I recognize that the taste is (sort of) the same- the pill has not made the cauliflower taste like chocolate cake, after all- but at the same time my experience is so different now that I resist saying that cauliflower tastes the way it used to taste. There is in any event no reason to be cowed into supposing that my cauliflower experiences have some intrinsic properties behind, or in addition to, their various dispositional, reaction-provoking properties. …
>
> After wearing inverting spectacles [spectacles with turn your visual field upside down] for several days subjects make an astonishingly successful adaptation. Suppose we pressed on them this question: "Does your adaptation consist in your reinverting your visual field or in your turning the rest of your mind upside-down in a host of compensations?" If they demur, may we insist that there has to be a right answer, even if they cannot say with any confidence which it is? …
>
> It is familiarly said that beer, for example, is an acquired taste; one gradually trains oneself – or just comes - to enjoy that flavour. What flavour? The flavour of the first sip? No one could like *that* flavour, an experienced beer drinker might retort: Beer tastes different to the experienced beer drinker … If we let this speech pass, we must admit that beer is not an acquired taste. No one comes to enjoy the way the first sip tasted. Instead, prolonged beer drinking leads people to experience a taste they enjoy, but precisely their enjoying the taste guarantees that it is not the taste they first experience.'

So Dennett is able to make a case for saying that there is no such thing as the properties that *our* experiences have, independently of how we choose to react to them. However, it is not clear that this fourth objection can succeed against the Knowledge Argument. The original argument requires only that we accept that Mary 'learns something new' when she leaves her room. Qualia are introduced to make it easier to describe *what* it is that she learns, but we don't have to use them to make the argument work. Suppose you think that talk of qualia is theoretically confused and so we shouldn't use them in philosophical theories. Nevertheless, it seems plausible that Mary has a new kind of experience when she leaves her room for the first time and sees coloured things, *and* that as a result of that experience she gains new knowledge about the world: she knows *more* about the colour-experiences of herself and other people. If you're willing to accept that, then the Knowledge Argument can go through even without making any mention of qualia.

Frank Jackson's change of heart

Frank Jackson has very publicly changed his mind about the Knowledge Argument; he now says that Mary does *not* learn new facts on leaving the room. Much of his reason for saying this is that he thinks that qualia, if they exist, must be physical, because qualia *do* cause things to happen in the physical world. One very straighforward example is that it is Jackson's experience of qualia that caused him to write about qualia. If qualia were epiphenomenal, as he originally claimed, then they could be part of a causal explanation of anything in the world, which would mean that qualia could not cause people's discussions about qualia! This,

he thinks, is absurd. Moreover, he says that the causal explanation for our having particular experiences will be a 'purely physical' explanation – so we *should* be able to deduce everything about our experience from the purely physical facts:

> Why do I think that the sensory side of psychology, as it is constituted in our world, is deducible in principle from enough about the world's physical nature? Our knowledge of the sensory side of psychology has a causal source. Seeing red and feeling pain impact on us, leaving a memory trace which sustains our knowledge of what it is like to see red and feel pain on the many occasions where we are neither seeing red nor feeling pain. … This places a constraint on our best opinion about the nature of our sensory states: we had better not have opinions about their nature which cannot be justified by what we know about the causal origin of those opinions.
>
> Now the precise connection between causal origin and rational opinion is complex, but for present purposes the following rough maxim will serve: do not have opinions that outrun what is required by the best theory of these opinions' causal origins. … We know that our knowledge of what it is like to see red and feel pain has purely physical causes. We know, for example, that Mary's transition from not knowing what it is like to see red to knowing what it is like to see red will have a causal explanation in purely physical terms. It follows, by the maxim, that what she learns had better not outrun how things are physically. …
>
> I now think that the puzzle posed by the knowledge argument is to explain why we have such a strong intuition that Mary learns something about how things are that outruns what can deduced from the physical account of how things are.' Jackson, *Postscript on Qualia*

Exam tip: You should notice that *each* of the three arguments covered recently (Chalmers' 'Explanatory Argument' deriving from the **Hard problem of Consciousness**, the '**Zombies**' **Argument**, and the **'Mary' / Knowledge Argument**) can be used not only to argue *for* Property Dualism; they also constitute arguments *against* any **physicalist** theory, and against **functionalism** and **behaviourism**. It's easy to see how an argument for property dualism is an argument against physicalism: the conclusion of each argument entails that there are some facts which are not determined by the physical facts, and thus that the subject matter of those facts must be non-physical.

The connection with anti-functionalist arguments is harder to understand. You have to add the extra premise that *a physical duplicate is also a functional and behavioural duplicate*: i.e. anything with exactly my physical makeup would also replicate my behaviour and the functional organization of my brain. That seems like a reasonable premise: how could something be an exact physical duplicate of me and yet have its internal workings behave in an entirely different way? Now we can say that, if it is possible for there to be a physical duplicate of me that lacked qualia, then it is possible for there to be a functional duplicate of me that lacked qualia, and so functionalism doesn't explain why we have qualia. Similarly for the Hard Problem. Functionalists explain the working of the mind in terms of the performance of functions; but what we really want to know (and an account in terms of functions cannot explain) is, why is the performance of these functions accompanied by conscious experience? In terms of the Knowledge Argument, we can say: Mary knows everything about the physical world, so she knows everything about the processes and functions performed by the human brain, but she doesn't know the facts about qualia: so facts about qualia are not facts about processes and functions, so functionalism cannot explain the whole of the mind.

Week 15: Consolidation

Read David Chalmers' article *Consciousness and its Place in Nature,* and make a note of the answers to the following questions (you should skip the section on 'The Two-Dimensional Argument...'):

1) Which of the philosophical theories you have studied should be counted under the headings of

- Type-A Materialism
- Type-B Materialism
- Type-C Materialism
- Type-D Dualism
- Type-E Dualism

2) What is the difference between an 'epistemic gap' and an 'ontological gap'?

3) What are 'psychophysical laws'?

4) What is 'neutral monism', and do you think it is a good view? Why?

Week 16: Extension material: Intentionality and the 'Chinese Room'

Intentionality

As we saw earlier, **intentionality** is the property of 'aboutness' or 'directedness' that some mental states have: for example, Riley's love for Buffy is a mental state that is 'about' or directed towards, Buffy herself. In fact, **propositional attitudes** in general (belief, knowledge etc.) are usually classed as **intentional states**, since these are all attitudes *towards* a particular proposition. Another thing that propositional attitudes have in common is that they are **representational** – they are attitudes towards a **representation** of the world as being a certain determinate way. Of course, the representation might not be one that I endorse: if am doubtful about whether Paris is in France, the object of my doubt is still a *representation*, just not one I currently endorse. But how can a materialist explain these mysterious qualities – intentionality and representationality – that so many of my mental states have? After all, we don't usually think of physical objects exhibiting intentionality – or if they do, it is because they are created with the purpose of representing the beliefs of a human being, as in the case of (for example) the text printed on this piece of paper. It certainly has intentionality, but only in the 'derivative' sense that it represents thoughts conceived in a human mind.

One answer to this is straightforward: the functionalist can claim that intentional states gain their representative power via their causal relationships. What makes my thought a thought about *dogs* is that my thought is appropriately related to my other beliefs about dogs. Of course, a belief also has causal relationships which connect it to *actions*, and the functionalist will want to say that this is also part of what identifies the thought as one about dogs: part of what makes this thought a thought about dogs is that it makes a direct difference to how I act towards *dogs*, and not much of a difference to how I act towards other people.

A problem for this is that it does not seem *sufficient* for my thought to be a thought about dogs that it has the right 'structural' relationships to other thoughts, to sensory inputs, and to behavioural outputs, as the functionalist suggests. Imagine a neuroscientist who came to know absolutely everything about the structural relationship between different elements of my brain. Would such a person be able to tell that my thought was about dogs? He might know that it had certain relationships to certain other thoughts, but that wouldn't be enough for him to work out that this was a thought about dogs *unless* he already knew that these other thoughts are also thoughts about dogs – and this seems to be something that cannot be 'read off' from the functional organization of the brain alone. (Further reading: Laurence Bonjour, *What's it like to be Human (instead of a Bat?)* Online at http://faculty.washington.edu/bonjour/Unpublished%20articles/MARTIAN.html)

Perhaps a more plausible approach is to claim that our decision to treat a system – any system – *as if* it possesses intentional states such as beliefs and desires is purely **pragmatic**, justified in its success in explaining and predicting that system's behaviour. This is the approach taken by Daniel **Dennett**, who uses the phrase 'the **intentional stance**' to indicate the attitude of treating a system *as if* it possesses intentionality:

> 'Here's how it [the intentional stance] works: first you treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not

all – instances yield a decision about what the agent ought to do; that is what you predict the agent will do.' Dennett, *The Intentional Stance* (1989)

Crucially, for Dennett there is no such thing as 'genuine' intentionality to contrast with 'as if' intentionality – the only thing we have to decide is whether it makes pragmatic sense to treat a given system (human or otherwise) *as if* it possesses intentionality.

Further Reading: Dennett, *Intentional Systems Theory*, online at http://files.meetup.com/12763/intentionalsystems.pdf.

The Chinese Room

John **Searle** offered a well-known argument against **Artificial Intelligence**, and against **functionalism**, claiming that the mere ability to mainpulate symbols (as a computer does), performing 'computational operations on formally specified elements', is not *sufficient* for understanding, and thus not sufficient for genuine intelligence; moreover, it is not sufficient for intentionality: since the computer does not understand the **meaning** of the formally specified elements, it is not capable of producing an output that is *about* anything in the same way that human speech is. Searle's own view is that 'only something that has the same causal powers as brains can have intentionality... whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origin as lactation, photosynthesis, or any other biological phenomena.'

Searle's argument is known as the **Chinese room**: read his 1980 article *Minds, Brains, and Programs* (online at http://www.class.uh.edu/phil/garson/MindsBrainsandPrograms.pdf), and answer these questions:

1) Make a diagram of the 'chinese room'!
2) In what way is the operator in the room like a computer?
3) What does the chinese room demonstrate about understanding?
4) What does Searle say is the relevant difference between his manipulating English and manipulating Chinese?
5) (extension work) Explain the following replies (and if possible Searle's responses to them):
    - the systems reply
    - the robot reply
    - the brain simulator reply
    - the combination reply
    - the other minds reply
    - the many mansions reply

Further viewing:

Searle - 'You can't do biology with beercans' at http://www.youtube.com/watch?v=KFiQX1qKjgQ.

Dennett on AI at http://www.youtube.com/watch?v=reMDjVM5ZiE.

Week 17: Extension material: Emergence and Biological Naturalism

The idea of **emergence**, or **emergent properties**, has been around for quite a while; it was originally used in the early 20th century to try to explain how the phenomenon of *life* could 'emerge' from more basic physical and chemical processes. Here's Samuel Alexander, writing in 1920:

> 'The higher quality emerges from the lower level of existence and has its roots therein, but it emerges therefrom, and it does not belong to that level, but constitutes its possessor a new order of existent with its special laws of behaviour. The existence of emergent qualities thus described is something to be noted, as some would say, under the compulsion of brute empirical fact'
>
> Alexander, *Space, Time, and Deity*

Emergent properties are 'higher-level' properties that are the *result of* or *emerge from* the organization of matter at the fundamental level, but which obey their own causal laws, and cannot be (in Searle's words) 'deduced or figured out or calculated' from objects at the fundamental level 'just by the way these are composed and arranged'. If mental properties are emergent, that might form part of the explanation of why they cannot be **reduced** to properties at the fundamental physical level, since the presence of an emergent property cannot be explained simply by saying how objects are arranged at the fundamental level.

One problem for theories of emergence is that it is very hard to find uncontroversial examples of genuinely emergent properties, and some philosophers doubt whether there is such a thing as emergence at all. For example, it used to be thought that *life* was an emergent property, but now scientists do not think that the phenomenon of life is mysterious at all – all there is to explain are processes such as metabolism, growth, and reproduction – all of which can be explained scientifically without the need for an emergent property of 'life'. Similarly, you might think that the view of consciousness as an 'emergent' property of biological systems is only a record of our ignorance of the processes that give rise to it, and that once these processes are found we will no longer think of mental properties as emerging in this mysterious way.

Another problem for emergence is the need to say *how* emergent properties emerge: is it just by magic? Searle offers an answer: for him an emergent process is one that is 'explained in terms of the causal interactions among the [fundamental] elements'. So, for example,

> 'The existence of consciousness can be explained by the causal interactions between elements of the brain at the micro level, but consciousness cannot itself be deduced or calculated from the sheer physical structure of the neurons without some additional account of the causal relations between them.' Searle, *The Rediscovery of the Mind*

So according to Searle, emergence is the result of 'causal interactions' at the fundamental physical level.

For Discussion: Does this provide a satisfactory explanation of how emergence works?

Biological Naturalism

John **Searle** is a **physicalist** who wants to avoid the problems associated with **property dualism**. Of course, if you want to reject the property dualist view that mental properties are a special kind of non-physical property, you have to explain how these mental properties somehow fit in to the physical world. Searle's answer is that mental properties are really a 'high-level' kind of **biological property** – part of the reason why he is so keen to reject **functionalism** is that he is convinced that not just any old system could have mental properties, consciousness, qualia etc.; these properties must in some way be the *result* of the way the brain is organized as a biological entity. For that reason, he has named his own view **biological naturalism**:

> 'To have a name, I have baptized this view, Biological Naturalism. 'Biological' because it emphasizes that the right level to account for the very existence of consciousness is the biological level. Consciousness is a biological phenomenon common to humans, and higher animals. We do not know how far down the phylogenetic scale it goes but we know that the processes that produce it are neuronal processes in the brain. 'Naturalism' because consciousness is part of the natural world along with other biological phenomena such as photosynthesis, digestion or mitosis, and the explanatory apparatus we need to explain it we need anyway to explain other parts of nature. Sometimes philosophers talk about naturalizing consciousness and intentionality, but by 'naturalizing' they usually mean denying the first person or subjective ontology of consciousness. On my view, consciousness does not need naturalizing: It already is part of nature and it is part of nature as the subjective, qualitative biological part.'
>
> Searle, *Biological Naturalism* (2004)

Searle is (broadly speaking) **anti-reductionist** – he rejects the idea that mental features such as consciousness can be **ontologically reduced** to count as nothing more than the arrangement of fundamental physical particles – and he accepts a form of **emergence**. However, he accepts a form of what he calls **causal reduction**, saying that the 'causal powers' of the mind can be explained solely in virtue of the causal powers of the biological structures that make it up. That's why he's able to insist that the mind is a **biological phenomenon** despite rejecting most forms of reductionism:

> 'The property dualist and I are in agreement that consciousness is ontologically irreducible. The key points of disagreement are that I insist that from everything we know about the brain, consciousness is causally reducible to brain processes; and for that reason I deny that the ontological irreducibility of consciousness implies that consciousness is something 'over and above', something distinct from, its neurobiological base. No, causally speaking, there is nothing there, except the neurobiology, which has a higher level feature of consciousness. In a similar way there is nothing in the car engine except molecules, which have such higher level features as the solidity of the cylinder block, the shape of the piston, the firing of the spark plug, etc. 'Consciousness' does not name a distinct, separate phenomenon, something over and above its neurobiological base, rather it names a state that the neurobiological system can be in.'
>
> Searle, *Why I am not a Property Dualist* (2002)

However, Searle's view has been criticized, especially by functionalists such as Jerry **Fodor**. Functionalists believe that minds can be 'realized' in *any* substance, provided it has the right kind of **functional organization**; they ask why Searle has a right to be so sure that consciousness must be a *biological* phenomenon which could

not in principle be 'realized' in other materials. Support for this criticism comes from examples in medicine: once the function of a biological organ is fully understood, that function can be replicated by mechanical systems with the right kind of functional organization. For example, our understanding of the human kidney has enabled us to construct mechanical dialysis machines which do all the same work as a natural kidney. Why couldn't the brain be the same? Surely once we fully understand how it works it will prove possible to construct a mechanical version which performs all the same functions – including, we might hope – conscious thought.

Further reading: Searle has a collection of online papers at http://ist-socrates.berkeley.edu/~jsearle/ (navigate to 'articles').

The Threat of Causal Inefficacy

A problem for many non-reductive physicalist accounts is that they risk treating the mind as **causally inert** or **causally inefficacious**. Suppose we take seriously the claim that mental properties are part of a 'higher-level' system which cannot be **reduced** to lower-level physical states and process. Nevertheless, your mental properties **supervene** on your physical properties. Does it make any sense to talk of **causation** at the higher, 'mental' level? Because supervenience is true (the mental depends on the physical), any change in your mental state must be accompanied by a change in your physical state – but because of the **causal closure of the physical**, every change in your physical state has a complete, sufficient, *physical* cause. This seems to suggest that events at the 'mental' level cannot cause events at the physical level – e.g. that my decision (mental event) is not the *real* cause of the physical event of my going to get more coffee, since this physical event *already* has a complete physical cause – no separate mental event is needed to make it happen; it would have happened anyway, even if there hadn't been a mental event. If the **causal closure of the physical** is true, it seems that all the genuine causation happens at the fundamental physical level, and the higher-level mental states, events, and processes are simply the *result* of what is going on at the fundamental level.

This is a serious objection because it threatens to undermine **rationality**. If we are rational beings, we must in some sense do things *because of* our reasons. But if everything we do can be fully explained by events and processes at the fundamental physical level, then it seems like our reasons are irrelevant to the real explanation of why we act as we do: the real explanation will involve neuronal activity in the brain, not the interaction of higher-level mental events. One possible response might be to argue that our actions are **causally overdetermined** – i.e. they are caused *both* by neuronal activity *and* by mental events such as decisions. But this is hard to make sense of: how can one event have two distinct 'complete' causes?