Philosophy of Mind
Excerpts from a Course Outline for lectures given at Sheffield, 1998

Peter Smith

Structure 2
1 Belief-Desire Psychology 3
2 Dualism, For and Against 7
3 Metaphysical Interlude 11
4 Perception 17
5 Belief 22
6 Levels of Description 25
7 Minds and Machines 30
8 Functionalism 33
9 Realism, Instrumentalism, Eliminativism 39
10 Sensations and Consciousness 43
11 Agency and freedom 52
12 Thinking and the Language of Thought 63
13 Representational Content 69

Version: January 1998
Course Structure

This Outline

This Outline is a very mixed bag—sections vary greatly both in the level of detail and in their relationship with Smith and Jones *The Philosophy of Mind* (sometimes the notes are just headline summaries of the book, but other sections add new material, amplifying or criticizing S&J, while others sections again go beyond anything in the book).

The length and detail of a section is not necessarily any indication of relative importance: brevity might just show that I haven’t upgraded headline notes into a more worked-through form.

§§ 12 and 13 (relevant to past special topics) are included for interest: they may or may not be relevant to this year’s special topics C and D.

Part of the notes for Special Topic B are based on teaching material kindly passed on to me by George Botterill. And large portions of some other sections of this Outline are closely based on a version written by Rosanna Keefe who taught this module in 1995 when I was on study leave. Very many thanks to her.

Sources

There is a list of reading given at the end of each section of this Outline. The philosophy of mind is an area well supplied with reprint collections of key articles, so there are often multiple sources for particular items on the reading list. Items marked [O] in reading lists should also be available in the off-print collection of the Short Loan Library. (It will be worth checking to see whether other papers listed become available in the off-print collection, which is being expanded.)

As already indicated, S&J will be used extensively in the course—both for the introductory section and as background for special topics A and B. It should be strongly emphasized, however, that you must also read widely from the topic-specific reading lists, and that familiarity with S&J alone will not be sufficient for any component of the course. On the other hand, some lists of ‘further reading’ are quite long (partly in recognition of availability problems), and it certainly isn’t expected that anyone read everything.

The following texts are also particularly useful and accessible introductions to parts of the subject:


An especially good collection of articles (even worth buying) is:


For an excellent overview of many of the topics we will be discussing (also no doubt useful for revision) see

1. Belief/Desire Psychology

1. States of mind

Let’s say something is minded if it has beliefs and desires, perceives the world around it, can act purposively in the world, experiences sensations, etc. We could say it ‘has a mind’ – as long as we don’t automatically assume that having a mind is a matter of having a special sort of component or part. (Compare ‘having a liver’ with ‘having a sense of humour’: the latter is a matter of being able to react to jokes etc., a question of capacities, not a matter of having a special organ. Which is ‘having a mind’ like? We mustn’t assume from the outset that it is like the ‘liver’ case.)

Contrast propositional attitudes reported using ‘that’ clauses (X sees, believes, desires, hopes, fears, intends, expects, wishes, tries to make it that case that p) with other mental states (particularly sensations—e.g. being in pain, having an itch). A theory of mindedness must have a story about both types of states, and the relationship between them (can a creature have sensations but no propositional attitudes? or vice versa?). But for the moment we concentrate on the first type.

Can we impose order on the multitude of propositional attitudes?

Perception involves acquisition of beliefs
Action is brought about by desires in the light of beliefs
Thinking is moving from belief to belief
Conscious awareness involves beliefs about one’s beliefs and desires

Here we use ‘belief’ as a rather wide, catch-all, for believing/taking-it-that/registering/thinking p: ‘desire’ is an equally wide catch-all for wanting/wishing/having the goal/lusting after etc. These seem central. That is, our basic folk picture is of mindedness as centrally a matter of having a belief/desire psychology: we register information which we use to secure our dominant desires. Many other propositional attitudes seem definable in terms of beliefs and desires – e.g. (roughly) X regrets that p if and only if X believes that p and desires that not-p—or at least involve beliefs and desires (as when X is sad that p, which requires at least believing that p).

2. Some questions that arise:

Can being a sufficiently complex organism suffice for having a belief/desire psychology, or does it require having a non-physical ‘soul’ or ‘spirit’?
Can states of believing this, desiring that, be states of the brain?
Can machines have beliefs and desires and be minded?
Is perception really just a matter of belief-acquisition?
Could our folk idea that we have a belief/desire psychology actually be wrong?
What is to be said about states such as sensations?
Is consciousness just a matter of having beliefs about our beliefs, desires, sensations etc.?
What, in particular, is involved in conscious rational thought (that which arguably distinguishes from more simply minded creatures likes dogs)?

3. Naturalistic constraints

We have a great and rapidly growing amount of knowledge about the causal origins of our actions in events in the brain, (about the way brain events are caused and themselves produce resulting activity in our peripheral nervous system in turn causing bodily movements, etc.). But intuitively, mental happenings are also causes of actions—as when I raise the glass because I want to drink the beer, or I run because I think that a bear is chasing me. A weak and surely acceptable naturalistic constraint on theorizing about mental happenings is

(C1) Our account of mental happenings as causes of action should be consistent with the main outlines of what we believe about the way that events in the brain are responsible for bodily activity.

[Note: that’s stronger than (C0) which says that our account of mental happenings should be consistent with what is true about the way events in the brain are responsible for bodily activity—that’s trivial because the truths about mental happenings have got to be consistent
with all other truths, since truths must be consistent with each other. (C1) adds, in effect, the assumption that current science is broadly right about what goes on in the brain.

Scientific enquiry into the brain is guided by the working assumption that human beings are part of the natural order, whose inner workings are subject to the same causal laws as apply across the rest of the natural world. The same principles of (say) cell biology apply to human brain cells as to other, lowlier, cells; the component organic molecules in our cells are subject to the same causal laws as apply to the same molecules found in other stuffs, and so forth. And there is no need to invoke mysterious ‘vital spirits’ or ‘soul stuff’ to explain what is going with us biologically. This suggests we can fill out (C1) like this:

(C2) Our account of mental happenings should be consistent with the view that we are complex organisms, part of the natural world, subject to the usual natural laws, lacking any spooky non-natural parts, etc.

A further idea is strongly embedded in scientific practice, namely that the behaviour of cells (the way they function and interact) is due to the chemical and physical properties of the constituent molecules; and the chemical and physical properties of the organic molecules concerned are again, at bottom, due to the physical properties of the constituent elementary particles. Chemists and biochemists are interested in (relatively) macro-properties of groups of atoms; but there is an entrenched methodological assumption that these macro-properties are the way they are as a result of the way the underlying micro-laws relate the relevant micro-properties of the component atoms. One minimal way of formulating this thought is as a supervenience thesis:

(S) Two systems cannot differ chemically or biologically without differing physically.

In other words, if two systems are identical in all physical respects, are particle for particle physical duplicates, then they are identical in all chemical and biological respects too (the chemical properties, say, cannot ‘float free’ from the physical ones). If we think of psychological/mental properties as being part of the natural order, it is plausible to add:

(SP) Two systems cannot differ psychologically (cannot have different intrinsic mental properties) without differing physically.

(In fact, for complex reasons, this needs some qualification, or needs a lot built into ‘intrinsic’: but certainly, SP – the thesis of the supervenience of the psychological – should look initially plausible to anyone who takes a naturalistic line about psychology and is inclined to the basic supervenience thesis S). So this suggests a third naturalistic constraint:

(C3) Our account of mental happenings should be consistent with, and explain, what is true in (SP).

These three naturalistic constraints are of increasing strength. It would be possible e.g. to acknowledge the force of (C1) while rejecting supervenience theses and hence rejecting (C3). We’ll mostly be appealing to (C1) and (C2).

(C2), the denial of the existence of spook stuff, is more a less a straight denial of dualism (the idea that mind is an entity separate from the body). For extended discussion of why various putative arguments for dualism give no adequate reason to reject (C2), see S&J, Part I; and §2 below.

(SP) is one version of physicalism but that term is a slippery one (different philosophers use it to mean subtly but importantly different things). We should distinguish supervenience physicalism from the theses e.g.

We are made of nothing but physical stuff.

Everything is physical: all entities, properties, relations, and facts are those which are studied by physics or other physical sciences.

Every psychological (biological, chemical, etc.) event is identical to some physical event.

Every psychological (biological, chemical, etc.) event is ‘realized by’ (or ‘constituted by’) some physical event.

4. One way to be a naturalist

Our biological theory of behaviour tells us that many bodily movements are caused by events in the brain. Our common-sense everyday psychology tells us than many purposeful actions are caused by (changes in) our beliefs and desires.

Thus: Jack’s arm goes up because of muscle contractions triggered by events in the peripheral nervous system ... triggered by changes at the neural level. But also, Jack raises his
arm because he wants to vote for Jill and believes that he will do so if he raises his arm now. How do these stories, the physical and the psychological fit together?

Here’s a simple theory (that evidently conforms to (C2) and (C3)): perhaps the physical and the psychological stories are just two different ways of talking about the very same thing. Maybe the event of Jack’s arm going up and the event of his raising his arm are the same event; and the underlying neuronal causes and the psychological causes are the same events as each other, differently described. Likewise for other mind–body relations: as it might be, being in pain is the same state as having one’s C-fibres firing (i.e. ‘being in pain’ and ‘having one’s C fibres firing’ are just two different ways of picking out the same state).

This (Type) **Identity Theory** says we can make this identification between the mental and the neural – *every type of mental state is identical to a type of physical state*. So the state of believing Sheffield makes steel would be a type of brain-state: anyone with that belief would have in common that state of the brain. Such identity claims are supposed to be comparable to the claims that temperature is mean kinetic energy; lightning is electrical discharge; or sound is compression waves travelling through air.

The Identity Theory is sometimes also called **Reductive Materialism**. It is a materialist theory since the only type of substance is matter – the mind is the brain. It is reductive because mental states are declared to be reduced to states of the brain: mental states are nothing over and above brain states. In the reduction of one theory to another, the truths of the former are accounted for in terms of the more general (and more explanatory) latter theory. It explains how entities of the reduced theory (in this case mental states) correspond to those of the reducing theory (in this case brain states). [Compare the fact that light has been reduced to electro-magnetic radiation. NB we do not deny that light exists because of this reduction, rather we say it *just is* electro-magnetic radiation. Similarly, we need not deny that there are mental states just because they have been reduced to brain states.]

5. **Three Unsuccessful Objections to the Identity Theory**

1) Mental states could not be brain states because the expressions for each (e.g. ‘pain’ and ‘C fibres firing’) do not mean the same.

But this objection carries no weight. In the same sense, ‘temperature’ and ‘mean kinetic energy’ do not mean the same either. It is an empirical discovery that the two different expressions (arguably) pick out the same property, not something that follows from reflection on the significance of the words. Identity Theorists can similarly admit the difference in meaning between mental-state expressions and expressions for brain-states: they are offering an **ontological** thesis about what there is, not a **semantic** one about what mental state terms mean.

2) Someone can know they are in pain, or have a certain belief without knowing anything about the state of their brain.

But compare Arnold who knows he is drinking water but not that he is drinking H2O, though water *is* H2O. There can be true identity statements of the form ‘A is identical with B’ whilst someone knows that something is an A without knowing that it is a B. To know that you are in a mental state doesn’t require complete knowledge about what it is to be in that state.

3) Further arguments along similar lines give properties which allegedly either the mental state has and the physical state does not, or vice versa. By Leibniz’s Law (if a=b then a and b have the same properties), this would refute the identity claim. E.g. mental states, but not physical states, have semantic properties: mental states can have content and they can be related to one another by entailment relations: states of the brain cannot have these features. Or: physical states, but not mental states, have spatial locations—-we would not say that my belief that the sun is shining is located two centimetres behind my eyes, but we could say this of states of the brain.

But the Identity Theorist can bite the bullet. If the two types of state are identical, then brain states *can* have semantic properties and mental states *do* have spatial locations.

But it may sound strange to make statements ascribing these properties, but similarly it used to sound strange to say that light has a wavelength. The discovery of the identities may bring with it other surprising facts.

U. T. Place recommends that the Identity Theory be seen as a **scientific hypothesis**. If true, empirical investigations should reveal correlations between observations taken to establish
that someone is in a given mental state and observations establishing that they are in a given brain state. The identity theory explains the correlation by maintaining that it is the same state being discovered to obtain. Because of the different sets of relevant observations (e.g. from introspection and from examination of the brain), we usually do not recognize the identity. Similarly the identification of lightning with electrical discharge is not trivial because the former is detected by causal observation whilst the latter cannot be.

6. Objection from Multiple Realization

The most powerful objection to the Identity Theory (in its original version, at any rate) is that mental states are multiply realizable.

1 Martians with silicon brains are conceivable: the Identity Theory will have to claim that they could have none of the same mental states as us since their brains could never be in the same physical state as ours. That’s implausible. Cf. different computers running the same program: different hardware states can realize the same software state. Couldn’t different ‘hardware’ realize the same mental processing?

2 If you don’t like sci-fi arguments, consider animals with brains unlike ours and which thus could not be in the same brain state as us, surely they could still have pains, perceive a banana, want a drink, etc. Yet the Identity Theory has to deny that.

3 Could it be objected that animals don’t really have pains, perceive their environment, etc.? That won’t help, because the plasticity of the human brain (e.g. relocation of function after damage) shows that there could be several ways of realizing the same mental states even in humans.

A naive version of type-type Identity Theory is therefore unacceptable. But other versions of Identity Theory may do better. (a) Identify mental types with rather high-level structural physical types? Or (b) move to a token-token theory ...

6. Type-type vs. Token-token Identity

The distinction between tokens and types is the distinction between particular things (tokens) which are instances of general types and those general types themselves. E.g. the distinction between this particular token or instance of the word the, and the word-type which is exemplified many times on this page. Likewise my belief that Sheffield is a city is a different token belief from your belief that Sheffield is a city, but these are beliefs of the same type.

The early versions of mind-brain identity theories, the sort that we have been discussing so far, were Type-Type theories, maintaining that every mental type (of state or event) is identical to a physical type – e.g. pains are C-fibre firings. It is these which are vulnerable to the multiple realizability objection. According to the Token-Token identity thesis, every token mental state is identical to a token physical state, (i.e. an instance of that physical type). But your token belief and my token belief, though beliefs of the same type, may not be physical states of the same type. Your pain may be C-fibre firing. Martian pain may involve a different physiology. More on token-token theories §8.

Reading:

Make a start on S&J, Part II. See also Crane MM: Chapters 1 and 2; Davies PM.

On naturalism, see


For basic reading on the mind-brain identity theory

Churchland MC: Chapter 2.3.


The classic papers on the identity theory are


2. Dualism: For and Against

1. Introducing Dualism

The Cartesian dualist (i.e. someone who holds the kind of position defended by Descartes) maintains that there are two radically different types of things: Minds (sometimes called ‘spirits’) and bodies which are hunks of matter.

Minds are essentially things which think, or are conscious, and are immaterial.

Matter is essentially extended—it has a position in space and time.

Minds are capable of existing without bodies and are not located inside them (since according to Descartes’ official line, they are not located in physical space at all). Persons in their normal embodied state consist of quite different components (two of them, hence ‘dualism’) somehow yoked together, a material body and an immaterial Mind.

NB our local convention here: ‘mind’ (small ‘m’) we use in the ordinary, pre-theoretical, everyday sense, in which it is uncontroversial that we have minds—i.e. are ‘minded’, i.e. have beliefs and desires and thoughts and intentions and sensations and perceptual experiences, etc. etc. By contrast ‘Mind’ (capital ‘M’) we use for the special sort of thing which the Cartesian postulates allegedly to explain why we have minds (ordinary sense). The existence of Minds is a controversial theory about minds: denying Cartesian dualism is not denying that we are minded in the everyday sense.

The Cartesian view of the mind is a package with other ingredients, strictly independent of the basic claim that mind and body are quite distinct entities. For example, it is claimed that we have privileged access to our own Mind. We can directly apprehend our own mental states and activities by a process of ‘inner observation’, known as introspection. Introspection is often supposed to be incorrigible—i.e. its deliverances are always right. E.g. we can’t be wrong about whether we are in pain. The mental is exactly as it appears to us to be. However, here we are going to concentrate on the basic dualist doctrine.

Dualism can be contrasted with both materialism and idealism. Materialists claim that the only sort of stuff is material stuff (humans are complicated material things, to be sure, but there is no spook-stuff in their make-up). Materialists need not deny that you and I are ‘minded’, have beliefs and desires and so on, but they will seek to account for this in material terms. Idealists hold that there is nothing except mind, claiming e.g. that physical objects are really collections of ideas (Berkeley).

As an alternative to Cartesian dualism some have defended bundle dualism (Hume’s view). According to Cartesian dualism, Minds are enduring things, independent objects in their own right. Mental states are states of those immaterial, objects and mental activities involve change in those states. According to bundle dualism, the mind is just a collection of mental events—the mind is nothing over and above the heap of pains, thoughts and so on that go to make up our mental history. Bundle dualists have difficulties accounting for the unity of each mind. Also, can there be experiences without any bearers to ‘have’ them? Although bundle dualism is often mentioned in introductory reviews it is not a very attractive view: Cartesian dualism has some pre-theoretic attractions, however (for example, it allows for life after death) – so as is usual we concentrate on that.

2. Arguments for Dualism

a) ‘Thinking about Venice and weighing 12 stone are hugely different properties: so the mental and physical properties must be had by different things.’

But why think that very different properties must belong to different things? (A cathedral is old, built of stone, visited by many, holy ...) A purely physical thing can have a wide range of properties. And mental properties are also of very different types (compare sensations and beliefs). Does that mean e.g. that what feels and what thinks are different substances (so Minds have to be split in turn)?

b) ‘A physical object couldn’t possibly have the properties we attribute to minds. No purely physical object could have beliefs, or use language, or fall in love.’

Why not? We have seen some of the powers of computers: robots can negotiate down round a room by a primitive vision system, seeking to fulfil some task—a primitive belief/desire system? (If not, why not?—thumping the table and saying a system
made of purely physical stuff can’t have beliefs and desires isn’t an argument.) Stones and grass may not be capable of these things, but increase the complexity of the object to the extent of the brain, say, and it is unclear why it could not be a belief/desire system. The dualist must not beg the question against the materialist.

c) An argument from introspection. ‘Pain doesn’t feel like the activity of neurones. We can tell by introspection that our mental states are nothing like any physical states.’

Why should we be able to tell by introspection everything about our mental states? (No other forms of perception reveal everything about their objects.) Why should our faculty of inner observation reveal things exactly as they are? It may just be that pain is exactly what having one’s C fibres firing feels like.

d) Linguistic arguments. “Jack” doesn’t mean the same as “Jack’s body”. “I admire Jack” doesn’t mean “I admire Jack’s body”—different objects of admiration are in question.

But two expressions can have a different meaning whilst picking out the same thing (e.g. ‘The Prime Minister’ and ‘the Member for Sedgefield’). We can explain the fact that the two expressions occur in different acceptable sentences by understanding “Jack’s body” as often functioning like “Jack, so far as his body is concerned” [see S&J pp. 24–27].

Besides, how far can we trust linguistic evidence to reveal the nature of mind? Maybe ordinary language just encapsulates a stone-age metaphysics: perhaps we think there are Minds when there aren’t.

e) Argument from survival after death. The mind can survive the death and destruction of the body, clearly the body cannot.

What reasons have we for believing in life after death? Religious authority is hardly conclusive. (Many religious doctrines can be understood compatibly with the denial of dualism, anyway.)

f) The following argument is due to Descartes: I can feign that my body does not exist, but not that I do not exist. So, I cannot be identical to my body.

This rests on the principle that expressions denoting the same thing can be substituted for each other into sentences without a change of truth value. But in the context of reports of psychological states (e.g. beliefs or imaginings), the principle does not hold. (Lois Lane believes that Clark Kent is Clark Kent, but does not believe that Clark Kent is Superman) [See S&J Ch. 3].

g) ‘No non-dualist theory can account for the character of our subjective experiential states’ (i.e. can tell us about ‘what it is like, from the inside’ to feel a pain, see red, etc.).

We will return to this negative claim in Special Topic A. But does the dualist say anything positive and substantial about these states anyhow?

3. Some Preliminary Difficulties for Dualism

a) Emptiness of the dualists’ theory of mind. The dualist has told us very little about the nature of Mind-stuff: no detailed theory, just the thesis that it exists and is not material. This can be contrasted with the great advances in theories of the brain, cognitive neuroscience, etc.

b) Neural dependence of mental states and events. Scientific evidence has shown the extent to which aspects of the mind (ordinary sense!) depend crucially on the functioning of the brain. E.g. damage to the brain can impair memory or the ability to bring perceptual experience under concepts and recognize what one sees, etc. Just why is this so on the dualist theory? More generally, can the dualist explain why (apparently) there are never Minds without complex brains?

The dualist might reply that the brain affects and is affected by the Mind, but the Mind is nonetheless a wholly different entity. Still, the more we discover about the brain and how its detailed workings affect mental phenomena, the less persuasive are the motivations for taking this line. (See 4.ii below)
c) The Evolutionary problem. According to evolutionary theories, the human race has evolved from other less intelligent species. The dualist must say that at some point in that steady process a radically new type of thing (i.e. Mind) suddenly came into existence. What explanation of this can be given?

d) Problems individuating Minds. What is the criterion by which we can individuate a Mind through time? It cannot be spatial-temporal continuity (cf. the criterion for this being the same spoon as I used yesterday – there is a continuous spatio-temporal path linking yesterday’s spoon to this one). And the ordinary first-person experience of continuity of the mind is neither necessary nor sufficient (not necessary, as I might suffer from amnesia but still be the same person; not sufficient since I might be deluded into thinking I am Guy Fawkes reborn). How can we distinguish the hypothesis that each ordinary person has one Mind though time from the idea that we have a new one every day, inheriting similar states to the old one? Likewise, how can we distinguish the hypothesis that each ordinary person has one Mind at a time from the hypothesis that they each have 17 Minds which normally run in parallel and ‘speak with one voice’? That idea seems not just empirically unlikely but some kind of absurdity, yet the dualist seems committed to thinking it makes sense. [See S&J pp. 46–49]


Apparently, the mental affects the physical: e.g. beliefs cause action; pain causes me to cry. And the physical affects the mental: e.g. physical damage to a body causes pain; images on the retina cause sensations (perception). Can the dualist accommodate this interaction?

i) ‘The mental and physical too different to interact!’ This would be a weak anti-dualist line to pursue. There is no requirement on causation that cause and effect be similar to each other. (E.g. smoking might cause cancer.) Take Hume’s account of causation as constant conjunction (roughly: c causes e if and only if events like c are always followed by events like e). This equally allows causal interactions between the mental and the physical.

ii) What sort of laws govern the interaction? Arguably all physical interactions can in principle be explained in terms of a few kinds of interactions of fundamental particles. By contrast, it seems that the dualist would be committed to a huge number of basic interactions between complex brain events and different types of Mental events. E.g. each different Mental sensation must be the causal result of the operation of a different psycho-physical law. [See S&J pp. 52–57.] Still, though that is an untidy result, it certainly is not fatal.

iii) A widely accepted scientific principle: the physical event is a causally closed system. That is to say, the chances of the occurrence of a physical event are fixed by earlier physical states (we need to talk of ‘chances’ here to allow for physical causation being non-deterministic). So, if we are looking for the causes of e.g. a neurological event, the alteration in the chemical and electrical properties of some brain-cell, then we need only look at earlier chemical and electrical events. Mental-physical causation, on the dualist story, would violate this principle. For suppose a physical event has an immaterial Mental cause. The physical state of the brain at the point at which the Cartesian Mental event causally kicks in will presumably not have had its chances fixed by the physical state just before. But there is no evidence at all for such a ‘gap’ in the chain of physical causation. Current science finds no reason at all to countenance the possibility. Bad news for dualism.

The dualist could accept that the physical event had a complete physical cause, but hold that there is widespread overdetermination. i.e. events have both complete physical and complete immaterial Mental causes. But this seems most implausible and the mental (as the dualist conceives it) becomes redundant for explanatory purposes.

Denying two-way interaction would open up two more alternatives within a dualist framework. But each has its own particular difficulties as well as sharing the other difficulties facing all dualist theories. Epiphenomenalism holds that mental events are caused by physical ones but can have no effects of their own in turn (mentality is like the froth on the wave). The case for there being mental events at all on this view is severely weakened: why suppose there are such things (construed as events in an immaterial substance) if they do no causal work? Parallelism holds that the mental and the physical run in parallel without either affect-
ing the other. Why? The undoubted correlations between the mental and the physical needs explaining.

None of the above difficulties for dualism is altogether conclusive, dualism can still be (and is still) maintained. But we have found no strong reason yet to believe in it, and in the light of our current scientific knowledge it looks considerably less plausible than at first glance. The score line looks bad for dualism (see S&J, Chapter V who compare it to flat-earthism).

Reading


Other introductory reading:
  Churchland MC: Chapter 2.1.

For Descartes’ own views, see:
  and *Discourse on Method*, part 5.

For a classic attack:

For latter-day dualists, see:
3. Metaphysical Interlude

1. Facts, properties, particulars

We need to grasp basic distinctions, between

**particulars**—items to which properties and relations may be attributed (and which are not themselves attributed to things). E.g.
- individual ‘concrete’ objects (the desk, this piece of paper, the Arts Tower, my car, your left hand, this carbon molecule …)
- abstract objects (the number seven, the set whose members are the prime numbers, a vector in Hilbert space, …)
- events, such as the stabbing of Caesar, the explosion of such-and-such a supernova, the moving of my little finger now.

**properties and relations**, which are attributed to particulars (the desk is *brown* and *dirty*; the desk is *in* the Arts Tower …). Note, we don’t confuse particulars with the linguistic expressions by which we pick them out—it is difficult to confuse the Arts Tower with the words ‘the Arts Tower’! It is as bad a mistake to confuse properties and relations with the linguistic expressions (‘predicates’) by means of which we pick them out. Being red or being in love are not to be confused with the predicates ‘…is red’ or ‘… is in love with …’. Properties and relations are the worldly correlates of predicates.

Some philosophers (‘nominalists’) have denied the existence of properties and relations, in this sense of the worldly correlates of predicates. On their view, what makes it true that this page is white is that this particular page resembles certain other sample particulars—all that exists in the world are particulars (and resemblances between particulars). The obvious snag is that resemblance looks like a relation (not a particular), so the world contains non-particulars after all. But we won’t pursue this here. The natural view is that properties and relations (*universals*) are as much a part of the furniture of the world as particulars.

**facts**—truth-makers for assertions, constituted (at least in the basic case) by particulars possessing certain properties or standing in certain relations.

Wittgenstein (in the *Tractatus*) says “The world is all that is the case” — i.e. what is ontologically basic is the totality of facts. On this picture, talk about particulars and about objects will be in some sense talk about what can be abstracted from facts. It is more common to think of facts as “built up” from independently existing objects and universals (a view not without difficulties). We needn’t worry too much about this dispute.

Rather than try to give sharper definitions, we’ll proceed largely by example (for our purposes, we want to minimize background metaphysical controversy—e.g. most will agree that there is some distinction to be drawn between particulars and their properties, even if they disagree about the best way to draw it, see below).

**Particulars**

Consider (a) the wrinkle in carpet, (b) the smile on Zoë’s face. Both are subjects of predication (have properties): the wrinkle is small but unsightly; the smile is warm and friendly. Are they particulars then?

We can say so. But intuitively, there is a sense in which these are not ‘things in their own right’, not ‘genuine individuals’, not ‘substances’ (Aristotle’s sense).

- Smooth the carpet, and the wrinkle ceases to exist; but it would be odd to say that any thing has been destroyed (cf. making the carpet cease to exist).
- The wrinkle depends for its existence on something else, namely the carpet; it cannot exist autonomously; its existence consists in the carpet having certain properties (its being arranged a certain way). Likewise, unlike her red coat, Zoë’s smile depends for its existence on something else; it cannot exist autonomously (cf. The Cheshire Cat vs his Grin).
- A carpet has (reasonably) clear criteria of identity across time: i.e. there is clear content to the question “is this carpet downstairs today the very same carpet as the carpet which was upstairs last year, or is it only an exact replica?” But “is this wrinkle here
downstairs today the very same wrinkle which was upstairs last year, or is it only an exact replica? looks a bogus question (what could the difference be?). A similar example: This morning Zoë wore a red coat and a smile; Yesterday Zoë wore a red coat and a smile. We can sensibly ask: “the very same red coat or just a replica?”; there doesn’t seem any content to “the very same smile or just a replica?”.

**Philosophy of mind examples (1)**

It’s pointless fighting about the label ‘particulars’. Let’s say that wrinkles and smiles are not basic particulars (they do not exist ‘in their own right’). What about e.g. ‘minds’ and ‘pains’ and ‘beliefs’?

These are subjects of predication (i.e. we attribute properties—Zoë’s stab of pain is intense but short-lived, her belief that her glass is empty is well-founded, and so on). But are they basic particulars?

**Descartes** held that minds are basic particulars. He distinguishes our material body—including the brain—from our immaterial mind or soul, which is an entity which exists in its own right and in principle independently of the body. (He is thereby committed to making sense of the distinction between there being the very same soul present on different days, and merely a replica. Kant argued that this commitment is problematic.)

NB denying Cartesian Dualism, denying that minds are basic particulars, is not denying that we are minded (have mental states); any more than denying that smiles are basic particulars involves denying that Zoë is smiling. For reasons for denying Cartesian Dualism, see these notes §2, and S&J.

**Hume** held that pains, beliefs, desires etc. are basic particulars (but minds, he thought, are not basic particulars but aggregates of them—this is so-called ‘bundle dualism’). NB To deny that e.g. pains are basic particulars is not to deny that we are often in pain.

**Events**

Should we treat events as basic particulars? They are subjects of predication, with properties. But, plausibly, the event of Zoë’s smiling at 10 pm is no more a basic particular than her smile is. (Could there be events which are basic particulars? Something which happens, but doesn’t happen to an autonomously existing particular? Perhaps. But the sort of macro-physical and psychological events we are concerned aren’t very plausible candidates for being basic particulars.)

But basic or not, they are an important class of particular—they feature crucially in many causal relations (it was the stone’s hitting the glass that caused the shattering of the glass).

It makes sense to identify events as being one and the same (the stone’s hitting the glass is the very same happening as the event that woke the baby). But what is the criterion of identity?

**Davidson**: Event E is one and the same happening as event F if E and F have exactly the same causes and effects (causally indiscriminability implies identity).

So: the event of Zoë’s thinking of Venice at 10pm could be one and the same event as certain neurological happenings *if* they have the same causes and effects.

If we think of events as (characteristically) non-basic, as a matter of independently existing objects changing in respect of their properties, then event identity will require identity of the objects and properties involved. So turn now to think about properties (universals) and *their* criteria of identity

**Universals**

Particulars exemplify universals—i.e. have properties and (together with other particulars) instantiate relations.

Proceeding by example again: putative properties might include having a mass of 2 kg, being red, being tall, having one’s C-fibres firing, thinking about Vienna. Relations might include: $x$ being taller than $y$, $x$ loving $y$, $x$ being between $y$ and $z$, $x$ being to $y$ as $z$ is to $w$, etc.

NB: Two different subject-terms can pick out the same particular, e.g.

‘Tony Blair’ and ‘The Prime Minister’
‘The Arts Tower’ and ‘The building with the longest paternoster in Britain’.
These are expressions with different sense (cognitive significance) but the same worldly reference. So likewise two different predicate expressions can attribute the same property, e.g. ‘is blue’ and ‘is Zoë’s favourite colour’

‘is at 43° Celsius’ and ‘has mean kinetic energy K’.

Again, these are expressions with different ‘sense’ but the same ‘reference’.

But when are the properties F and G identical? When do the expressions ‘F’ and ‘G’ pick out the same property? Not (as noted) just when ‘F’ and ‘G’ have the same sense.

One plausible suggestion (for empirical properties): F and G are the same property if they are causally indistinguishable: i.e. if causally bringing it about that x is F brings it about that x is G and vice versa, and if x’s being F gives x the same causal powers as x’s being G. For example: you make it the case that x’s being at 43° Celsius by giving x a certain kinetic energy (and vice versa); and being 43° Celsius gives x the same causal powers as having a certain kinetic energy. Similarly make something blue and you make it Zoë’s favourite colour, and things that are Zoë’s favourite colour reflect light in just the same way as things that are blue, etc.

(Why won’t it do just to say that F and G are the same property if x is F if and only if x is G? That’s too weak. Every x which has a heart has kidneys and versa versa, but intuitively having a heart and having kidneys are not the same property. Why won’t it do just to say that F and G are the same property if, as a matter of logical necessity, x is F if and only if x is G? That’s too strong. Being painted blue and being painted Zoë’s favourite colour may be having just the same property, but it isn’t a matter of logical necessity that Zoë’s favourite colour is blue—in some worlds, her tastes differ.)

Philosophy of mind examples (2)

What is the relation between the property of being in pain and the property having one’s C-fibres firing? Obviously the expressions ‘... is in pain’ and ‘... is having C-fibres firing’ do not have the same sense. But the suggested criterion for identity of properties allows room for the claim that the expressions pick out the same property.

And, on the suggested criterion, whether this is true will be a matter for empirical discovery. Is it the case that being in pain is co-instantiated with, and has the same causes and effects as, having your C-fibres firing? Likewise, is it the case that believing that this paper is white has the same causes and effects as having neurological property W (for some appropriate W)? The multiple realization objection to identifying mental and physical properties comes to the claim that the bringing it about that something believes this paper is white need not involve bringing it about that it is W.

Types vs. tokens, and events again

Take the question: how many words are there in the displayed sentence?

The cat is on the mat

We might say six; or on another principle of counting five (one word appears twice). In the first case, we are counting word tokens, particular inscriptions. In the second case, we are counting word types, i.e. counting the number of universals instantiated by the inscribed tokens. (Context should usually make it clear whether it is types or tokens, universals or particulars, that are in question.)

Likewise we should distinguish between type and token events. Someone utters that sentence; how many different events of word utterance? Five type events instantiated by six particular token events.

When we ask in the philosophy of mind whether, e.g., events of feeling pain can be identified with some neurophysiological happening, we need to be clear whether this a claim about particular events or types of events. Compare the templates

This particular token pain event is identical to such-and-such neural event

This type of mental happening (a pain event) is identical to such-and-such type of neural event.

The second, type-identity, claim might be refuted e.g. by mice or Martians who feel pain but have a different physiology; the first token-identity claim would not be.
Identity, constitution and realization

First, we’ll consider the notions of identity, constitution and realization as they apply to particulars.

We need to distinguish numerical from qualitative identity. Compare:

A. Jill and her twin sister Jane are (let’s pretend) completely identical. You and I have identical sweaters. Jack and Jim are wrapping up identical bottles of Chanel to give to their respective beloveds for Valentine’s Day.

B. Jill’s mother is identical to Jane’s (she is one and the same person). The sweaters were bought from the identical branch of Marks and Spencer. Jack’s beloved is identical to Jim’s—one and the same woman is the object of both men’s desires.

In type (A) cases, we have qualitative identity—Jill and Jane share the same properties. Likewise for the two sweaters, the two bottles of perfume. In the (B) cases, we have instances of numerical identity—it is one and the very same entity (mother, shop, woman) that is in question each time.

The basic principle governing numerical identity is Leibniz’s Law

LL Necessarily, if \( a \) and \( b \) are identical (i.e. are one and the same thing), then whatever property \( a \) has, \( b \) will have too.

E.g. since Jill’s mother is one and the very same person and Jane’s mother, whatever property is had by Jill’s mother must be had by Jane’s mother. If Jill’s mother is a logician, so is Jane’s; if Jill’s mother has three children, so does Jane. And so on.

Conversely: if \( a \) has some property that \( b \) lacks, then \( a \) is not one and same thing as \( b \).

Next, constitution. What is the relation between e.g. this statue \( S \) and the lump of bronze \( B \) which forms the statue?

Can we say ‘\( S \) is one and the same thing as \( B \)?’ Arguably No: for \( B \) has properties that \( S \) lacks. (E.g. \( S \) came into existence in 1998; the lump, let us suppose, had been hanging around waiting to be reshaped for years—likewise \( S \) goes out of existence in 1999 while the lump continues to exist, reshaped. So \( S \) and \( B \) satisfy different persistence conditions).

Can we say ‘\( S \) is one thing, \( B \) is another; they are not identical.’ But for now, \( S \) fills up exactly this region of space: how can \( B \) occupy the same space at the same time? Here’s one story how: An object is present at a time by having a temporal part that overlaps that time. \( S \) and \( B \) can share temporal parts. In fact the temporal parts of the statue are all identical with some temporal parts of the lump of bronze: but the lump also has temporal parts that are not temporal parts of the statue. The statue and the lump are thus sums of different collections of temporal parts (that’s why they are different); but the sums overlap, sharing some parts (that’s why \( S \) and \( B \) can be in the same place at the same time).

That looks neat. But can we really accept a ‘temporal parts’ metaphysics? Is it really the case, for example, that parts of you aren’t here now?

So should we say instead then that ‘\( S \) is constituted by \( B \), where constitution is neither identity nor overlap of (temporal) parts? NB it is uncontentious that \( S \) is constituted of bronze (that is a relation between the statue and a kind of stuff); what is contentious is that there is a distinctive ‘constitution’ relation between two objects, the statue \( S \) and the particular lump \( B \).

So what is the relation of constitution between objects?

Asymmetric (if \( A \) constitutes \( B \), then \( B \) does not constitute \( A \)—cf. identity);

If \( A \) constitutes \( B \), then \( A \)’s parts are parts of \( B \); but (???) example suggests a constituting object has part-dependent (‘mereological’) criteria of persistence. You track it through time by tracking its parts (at some relevant level). [The Lego model.]

Role-realization. Arguably, we sometimes use e.g. ‘The Prime Minister’ to refer to a particular person, sometimes to the role. Consider, compare

The Prime Minister visited Sheffield today.

The Prime Minister has the power to select the members of the Cabinet.

The latter is naturally taken, in context, as meaning: whoever fills the role has the power. And when we say ‘Tony Blair is the Prime Minister’ we may be reporting that he fills the role (or ‘realizes’ the role). Another example: ‘This is my partner’.

We want then, to distinguish

the ‘is’ of identity: Napoleon is Bonaparte
the ‘is’ of constitution: A human body is a collection of cells
the ‘is’ of role-filling: Zoë is his Valentine

**Philosophy of mind examples (3)**

Consider the relation between Zoë and her body.

**Zoë is identical with her body**

- But (1) Sci-fi cases of a person surviving in a different body (brain transplants, for example)?
- (2) Zoë can think about Venice, can do long division in her head; can Zoë’s body think about Venice, do long division? (If not, then apparently the person and her body have different properties so are different things).

**Zoë is constituted by her body**

- But (1) Bodies do not have part-dependent criteria of persistence (they survive continual replacement of cellular material). So need a more complex story about constitution.
- (2) ‘Zoë is a human being; human beings are not constituted by their bodies (organisms)—they simply are organisms.’ [So Zoë is identical with her body after all.]

**Zoë is a person-role realized by her body**

- But (1) Makes Zoë a type rather than a token; not a particular at all, let alone a basic particular.
- (2) As before: ‘Zoë is a human being; human beings are not roles realized by their bodies (organisms)—they simply are organisms.’

**Properties and states**

Turn to questions about identity, constitution and roles applied to properties and states.

Aside on states: Philosophers talk of e.g. Zoë’s believing that her glass is empty, or feeling toothache as mental ‘states’. What they intend is in part a contrast with fleeting events—so it is the possession of some mental property over some stretch of time. States will be identified via the property whose possession they consist in.

We’ve already given a criterion for property identity (and hence state identity) in terms of causal indiscriminability. (In a slogan, a difference that makes no difference is no difference at all.)

Can we talk about the constitution of states (and the properties they involve)? Perhaps: ‘The property of being a good all-round cricketer is constituted by the combination of the properties of being a good batsman, a good bowler and a good fielder’. But this just reminds us that there are complex, conjunctive, properties whose possession is entailed by the possession of some other properties. Are there cases of one state ‘constituting’ another which aren’t just cases of logical entailment? Well, we might say the state of having a gene for blue eyes is constituted by such-and-such a state of one’s DNA. But surely talk of realization is more apt here: ‘Being a gene-for-blue-eyes is a causal role: it is a matter of playing a certain part in inheritance of characteristics: it has turned out that having certain properties in our DNA realizes that role in us.’

**Philosophy of mind examples (4)**

What is the relation between being in pain and having one’s C-fibres firing (as it might be); or wanting a glass of beer and being in such-and-such neural state?

*The mental state is identical to the physical state.*

- But: multiple realizability objections if this is a type identity claim.

*The mental state is constituted by the physical state.*

- But: No clear notion of state constitution in play (the simple logical entailment sense won’t do)

*The mental state is fixed by causal role; the physical state realizes that role.*

Much more on this in discussing functionalism.
Reading

There are very useful chapters on metaphysics in

See also

D. H. Mellor and Alex Oliver, ‘Introduction’ in D. H. Mellor and Alex Oliver eds.
4. Perception

Perception: vision, hearing, taste, smell, touch and proprioception (or body sense: perception of the position of parts of one’s body).

Veridical perception involves appropriate causation: if I am really seeing that the cup is full of coffee then it must be the cup and the coffee which is causing that perceptual experience (not, e.g., a series of illusions which happen by accident to match the truth). Other perceptual-type experiences, such as illusions, hallucinations, after-images, are causally brought about in non-standard ways.

It is tempting to say that when I hear a distant train approaching, the train itself is not the immediate object of my experience. Rather my perception of the train is mediated by my hearing the sound it makes—i.e. I hear the train in virtue of hearing the sound. Let’s say that something is a mediate object of perception if it is perceived by perceiving something else (perhaps, to take another example, I see the train as it comes down the track by, in the first place, seeing its front side). An immediate object of perception is one which is not perceived through the perception of something else. Often non-immediate perception involves more or less conscious inference, the reliance on collateral information (e.g. you infer that there is a train approaching because you take the sound to be the sound of a train).

What are the immediate objects of experience? I see the train in virtue of seeing the front side: but it might be argued—I see the front side in turn in virtue of being aware of an array of coloured expanses of a certain size and shape, where these expanses are in my private visual field. Press my eyeballs and they go blurred, shut my eyes and they cease to exist: things in the world don’t themselves blur or go out of existence when I press or shut my eyes. So the coloured expanses that I am immediately aware of are private, mental data. This line of thought motivates...

SENSE-DATA THEORIES OF PERCEPTION

The sense-data theorist claims that the immediate objects of our experience are always private mental data (variously called ‘ideas’, ‘percepts’, ‘sensa’, ‘sense-data’ by the various British Empiricists from Locke and Hume to Russell and Broad). The properties of sense-data are those which are immediately presented to you, in how things look (or sound or feel, etc.). If it looks to me as if there is a grey elephant before me, then I must have a grey, elephant-shaped visual sense-datum. But the sense-data do not intrinsically carry the information that it is a visual experience of an elephant—this depends on the inference, based on background knowledge, that the sense-data are produced by an elephant.

Sense-datum theorists tend to defend the representative theory of perception [e.g. Locke, Frank Jackson]. There are external objects which our sense-data represent. For veridical perception, (some) properties of sense-data appropriately match properties of the thing in external world, and the occurrence of the sense-data is causally dependent on the world. With illusions, the sense-datum has properties which the external object does not have and hallucinations have no external object corresponding to the sense-datum.

Gives rise to a troublesome sceptical problem: If we can never perceive material objects directly (they are hidden behind the “veil of ideas”), what justification do we have in believing that they exist? One response to this epistemological problem is idealism – this retains the sense-datum theory while demoting ‘material’ objects to some kind of construction out of sense-data. Illusions/hallucinations yield sense-data which do not cohere with other sense-data. No comparable sceptical problem about how we know about the external world – there is no external world! Chairs and tables are complexes of ideas. (But what about the unperceived tree in the depths of the forest? “Ah,” says Berkeley. “God sees it ...!”) However, we won’t explore the epistemological implications of the sense-datum theory here.

Sense-data theories of perception of the various kinds are traditionally contrasted with Direct Realist theories, according to which the immediate object of experience can be a material object and the external world is thereby perceived directly, without the mediation of sense-data. What’s to chose?

The Argument from Illusions/Hallucinations
The main argument for sense-data. The claim is that the experience had when hallucinating an elephant can be exactly the same as the experience of seeing a real one. The sense-data theorist explains the similarity by maintaining that the immediate object of perception in these two cases is the same. In the hallucination, the subject is aware of something: the sense-data theorist claims that it is her sense-datum of which she is aware – in the visual case, an array in her visual field. Then it is argued that, since the experience in the veridical perceptual case is exactly the same, we need to account for veridical perception in terms of the awareness of sense-data too.

An opponent needs to provide an account of perception and illusion which allows the felt similarity which they can have, and yet which can account for the difference. The Direct Realist might be thought to have problems. In veridical perception it is allegedly the external objects that we see directly, but with hallucinations there are no such external objects—how can the experience be just the same?

Some Initial Questions

Is the sense-datum theory equally plausible for non-visual perception? (Compare “in seeing, we are aware of visual images” with “in proprioception, we are aware of ...”—well, what?). Can sense-data exist unperceived? (We assumed not above. But Russell maintained that they could. What sort of dispute is this?) Can sense-data have features of which we are unaware? (Or is there a determinate number of spots on the sense-datum of a speckled hen, even though I can’t tell how many?) If I look at the unchanging computer screen, shut my eyes, and look again, am I interruptedly aware of one sense-datum of a screen or two sense-data just the same? (Questions like this crowd in, all looking rather suspiciously artificial.) What is the ontological status of sense-data?—traditionally they were conceived within a dualist picture of the mind: do they inherit the problems of dualism?

The Fatal Regress?

If e.g. seeing involves having mental images, how do we perceive those images? And is this also done through other images? Unless we can answer these questions, the Sense-data theory has provided a very unilluminating account of how we perceive. Telling the awareness-of-images story about the perception of the internal images themselves threatens to need an unending series of little people in my head seeing images-of-images just to capture what is going on in my own perception. Alternatively, the sense-data could be images which can just be had without being perceived, but this claim is obscure as it stands: surely images work by being seen? (Cf. Dennett’s story of the man who riggs up a camera on the front of his car, and a TV screen under the bonnet “so that the car can see where it is going”: the images are useless without something to see them.)

THE BELIEF-ACQUISITION THEORY OF PERCEPTION

“Seeing is believing”. Seeing a pink elephant is acquiring the belief that there is a pink elephant in front of you. (NB, the seeing is to be identified with the acquisition, not what is acquired.)

Perception of external objects can be direct, i.e. not mediated by the awareness of internal mental objects. Thus problems with the existence and nature of such mental entities do not arise. For veridical perception, true beliefs are acquired (in the right causal way), whilst with illusions (and hallucinations) it is typically false beliefs that are acquired.

Perceptions represent the world, as do beliefs. But the world need not be how they represent it as being—beliefs need not even be about real things. So illusions and hallucinations need not be a problem for the theory. Illusions and veridical perception can be similar—in both beliefs are acquired. At one time a given belief could be acquired during an illusion, whilst at another time the same belief could be acquired from veridical perception. The veridical/illusion distinction can still be maintained by noting differences in whether or not the acquired belief is true at the time it is acquired.

This account is amenable to a functionalist picture of the mind, since perception is characterized by its typical effects: change of belief. Combined with a functionalist account of
beliefs as dispositions to behaviour, connections between perception and behaviour are established. This explains the ability to get around the world and to react to our environment which we perceivers have. Moreover, there is no threat of regress in the analysis comparable to the threat to sense-data theories.

Unconscious Perception and The Phenomenon of Blind-Sight

Can there be unconscious perception? Suppose when driving along I negotiate a log in my way by driving around it, but I am not aware of doing so (I am too busy chatting to my companion). I have perceived the log (that is how I managed to avoid it). We could say that I have acquired the corresponding belief, but that belief was not a conscious one.

‘Blind-sight’ describes the abilities of certain subjects who have no conscious visual experience of objects in a certain area of their visual field. They claim to be unable to see anything in that area. But they can e.g. catch a ball approaching them from that direction, and react to other things, revealing that they have, in some sense, acquired information about objects in those regions and the ability to discriminate things there (whilst all the time denying this). Are such subjects perceiving?

Such cases as blind-sight are problematic for sense-datum theorists (even when asked to concentrate, the subject reports no experience, no sense-datum). The belief-acquisition theorist just cheerfully says that the subject is perceiving and acquiring beliefs (since they are disposed to act in the right kinds of ways), but the corresponding beliefs are not available to consciousness.

Problems and Modifications

a) We talk about perceiving things (e.g. seeing cats and hearing clarinets) not just perceiving that something is the case (e.g. that the cat is on the mat or that the clarinet is out of tune). But beliefs must be beliefs that..., they can’t have cats and clarinets as their content.

The basic beliefs acquired in these cases are still beliefs about certain states of affairs – e.g. that the corresponding things are in the vicinity, over there, etc. …

b) Part of my perception may involve e.g. a small yellow blob seen out of the corner of my eye: do I really acquire a belief that such a blob exists?

Don’t read too much into talk of ‘beliefs’: to repeat, we don’t mean considered, conscious states – just unreflective takings-it-that. If I were in a laboratory experiment on peripheral vision, asked e.g. to press a button when I sensed a yellow blob, then I’d press the button, evidencing that I indeed took it that there was a blob.

c) When you see the white wall in front of you, you don’t always acquire the belief that the wall is white—often you knew (and believed) that already.

It is still true that if you had not already believed it, you would then have acquired that belief. (More subtly, it might be said that in such cases the information acquired through perception updates tensed information to the present – you acquire the belief that the wall is (still) white now.)

d) You can have illusions without being fooled by them. You need not acquire the corresponding belief that e.g. there is a pink patch of wall in front of you (you know it is just an after-image).

In these cases the tendency to believe is still there. You would acquire the corresponding belief if you didn’t have certain other beliefs which you have for independent reasons (e.g. that the wall is actually uniformly white).

But note, it is controversial whether this response is adequate. Cf. Jackson.

e) Sometimes the beliefs you come to acquire do not reflect what you perceive. E.g. you could come to believe that something is red though you see it as khaki-coloured, because you know that red things look that shade of khaki in the given sodium light. Must the theory say that you see it as red?
Again, the response can invoke a *tendency* to believe. If you had not had beliefs about the effects of the lighting conditions, you would then have believed it was khaki on the basis of the perception.

f) More needs to be said about the distinction between the cases when acquiring the belief that p is perceiving that p and when it is not (you can acquire beliefs e.g. just by reflection and inference).

Add that in perception the beliefs are acquired by means of our sense organs: the relevant sense organs were a causal factor in the belief acquisition.

What about proprioception? And when you read something telling you that p, the belief that p is also acquired by means of the senses but you are not perceiving that p.

The theorist could borrow from the mediate/immediate distinction. In reading that p, my belief-acquisition is mediated by acquiring beliefs about the disposition of letters on the page, etc. It is the non-mediated, non-inferred, belief-acquisitions which are perceptions. (Could we also add: seeing that p involves acquiring the belief that you are seeing that p as well the belief that p itself? Troublesome: animals can see, but do they have beliefs about their own mental states?)

The Fundamental Problem?

What is the belief I acquire on tasting something I do not recognize, and which I have no concept for? What is the belief I acquire when I see a particular shade of blue for which I have no prior concept? More generally, don’t the contents of our perceptions outstrip the conceptual contents of our beliefs?

Note, this isn’t exactly the objection that “a picture is worth a thousand words”, i.e. that we may have no words to capture the details of particular perceptual contents. For we can have unverbalized concepts (the dog recognizes a rabbit when he sees one, but has no words at all).

The worry is that when we do form beliefs on the basis of perception, we are ‘chunking-up’ the acquired information, pigeon-holing it (applying the concepts “red”, “rabbit” or whatever) – there is a loss of informational richness but with a corresponding gain in ease of processing. We see the traffic light (a particular precise shade of red): in bringing it under the coarse concept “red” we lose informational detail, but this judgement usefully now engages e.g. with the rule “stop at red lights” and we can now easily process the combination of rule and coarse-grained information and deduce “stop!”. Likewise, the dog brings its perceptual experience under the concept “rabbit”, and this combines with the standing belief e.g. that rabbits are good to eat, to initiate a chasing routine.

The thought, then, is that conceptualization is applying a useful coarse-graining to the richer perceptual information (Dretske: it is moving from analog to digital representations). So perceptual information is not the same as conceptualized information.

Still, it might be said that in some respects, this motivates a modest retreat rather than full-scale abandonment of the belief-acquisition theory. It should have been claimed, rather, that perception is the acquisition of information leaving it open whether the information is chunked-up conceptually. This is still a substantial, and contentious, thesis: for it implies that we exhaust the nature of perceptual experience by giving its representational/informational content. (There are various ways of filling in the space around me compatible with how things look—and the representational content of my experience might be identified with the set of these ways of filling in space compatible with the way things look. Does fixing this content fix everything about my experience?) It might be argued that perceptual experience also has ‘sensational’, non-representational aspects. Compare: some aspects of a painted picture are representational – that this blob is to the left of that represents the apple as being to the left of the orange. Other aspects, the brush-stokes perhaps, are not representational. Couldn’t it be like that with perception? (Consider the inverted spectrum story. You see as red what I see as green and vice versa. When we look at pillar boxes we get the same worldly information, i.e. that they reflect light like ripe apples and stop signs and unlike grass and unripe apples. But we see them differently. The non-representational aspects of our experience differ. Isn’t that possible?)

We leave the matter here, wondering whether informational theories of perception pay enough attention to the qualitative nature of our perceptual experience.
Reading

Basic reading

S&J: Chapters VII and VIII.
Chapters 10 & 11

Further reading

Especially chapters 2, 3 and 9.
Crane’s introduction and the papers by J.J. Valberg and E.J. Lowe.
5. Belief

Some Observations about Beliefs

1) Beliefs are mental states with content. The content of my belief that it is raining is the same as the content of Françoise’s belief which she expresses by the sentence “il pleut”—i.e. the content of my belief is not the English sentence which expresses it, but the proposition expressed both by “it is raining” and “il pleut”. Hence the talk of beliefs as ‘propositional attitudes’—in this case, the attitude is acceptance. Other propositional attitudes include desires, desires and fears: these too have propositional content. I can believe that p, while you desire that p, and she fears that p: we have different attitudes to the same proposition.

2) A subject’s system of beliefs can be seen as her ‘map’ of the world: how she takes it to be. The beliefs represent (but can misrepresent) the world. There is a sense in which they essentially ‘aim at the truth’. (Cf. Moore’s observation that it is absurd to say ‘I believe that p but p is not true’. Of course such a saying could be true: I can believe something and be wrong! The absurdity is in saying it, putting myself forward as believing that p when I am also in a position to assert that not-p).

3) Beliefs play a central role in guiding action. My belief that there is beer in the fridge will enable me to fulfil my desire for beer by going to the fridge. This indicates a close relationship between beliefs and behaviour (see below).

4) Beliefs are sometimes said to be occurrent or non-occurrent. The occurrent beliefs you have at a particular time are those your attention is on, or which you are thinking about, or are actively governing current behaviour (or some such). Contrast: you believed that there are no bright green giraffes before reading this, even though it was not then an occurrent belief.

5) Contrast ‘belief that p’ (where p is a proposition) with belief in something (e.g. a belief in God). Belief in something could plausibly be taken to reduce to belief in a collection of propositions (e.g. perhaps believing in God = believing that God exists; or belief in a person is often taken to be belief that they are responsible, trustworthy etc.)

6) Beliefs can come in degrees, in the sense that we can be more or less confident in the truth of some proposition.

7) Beliefs are in important respects involuntary. Sure, you can choose what to think about, knowing this will lead you to acquire new beliefs with a certain subject matter, but you are not then choosing exactly what you will come to believe.

8) Some beliefs tend to give rise to certain others to which they are inferentially related. E.g. if I believe that all ravens are black, I’m very likely to believe that the next raven I will see is black. If the first of these beliefs is true, it follows that the second one is also true, and it is rational to have the latter belief if I have the former. This is not to deny that we often fail to draw all the consequences of our beliefs (life is too short to do that!): it is a serious problem for designers of AI systems to get their systems to draw enough ‘beliefs’ from their information stores, while ignoring irrelevant inferential consequences.

Hume’s Account of Belief

Hume claimed that having a belief is a matter of having an idea or image in the mind and having a certain feeling of conviction associated with it. (Beliefs are taken to be self-intimating, i.e. if you believe that p then you know that you believe that p.)

This account will at best only work for occurrent/conscious beliefs. We certainly do not have occurrent feelings corresponding to non-occurrent beliefs. The suggestion that we have an image for each of our beliefs (even just the occurrent ones) is implausible. There is no particular image that goes with my belief, say, that Spanish would be easier for me to learn than Chinese. Moreover, when there are potentially appropriate images in the offing, we can have
them in mind without having the relevant belief, in particular if we merely ‘entertain the proposition’. Hume claims that the difference is that with beliefs the image is more vivid and forceful. This is a highly implausible account. I can have vivid images (e.g. in imaginations) without an accompanying belief. So having a vivid image is neither necessary nor sufficient for having a belief—Hume’s account must fail. It also cannot explain the close relationship between beliefs and behaviour (nor why some beliefs tend to give rise to certain other beliefs or to exclude various others.)

**Beliefs and Behaviour**

Beliefs are arguably not merely contingently connected to behaviour (to suppose that someone could believe that the cup contains coffee, but in every way be disposed to act as if it contained cyanide seems some kind of absurdity). That is to say, it is part of what it is to have a given belief that it tends to give rise to appropriate behaviour, in particular behaviour that—were it and the subject’s other beliefs true—would lead to satisfaction of the subject’s desires. The belief that the coffee is in room 12.12 leads me to go there, given that I want coffee—and I’ll get coffee, my desire will be satisfied, if my belief is true.

This suggests using this connection with behaviour to characterize beliefs. This may provide a better account of the difference between believing that it is about to rain and simply entertaining the proposition—the latter does not in general lead to action. Also since non-occurrence beliefs can be connected to behaviour, the limitation of beliefs to conscious episodes can be avoided.

**Behaviourist accounts**

Reductive behaviourism claims that talk about mental states such as beliefs just is disguised talk about actual or potential patterns of behaviour (belief-talk can be reduced to behaviour-talk). When you say that Jack believes that the cup contains coffee, you are just saying that Jack is disposed to suitable behaviour. (We’ll consider here just behaviourism about intentional states like beliefs; full-blooded behaviourism holds that sensations too are just dispositions to behave—e.g. when you say Jill is in pain you are saying that she is exhibiting or is disposed to exhibit pain behaviour. For example, that she is disposed to wincing and clenching her fist etc. Behaviourism seems distinctly less plausible for experiences with a ‘felt inner quality’, as we are tempted to put it. What about the paralytic or perfect actor who is in pain, but exhibits no tendency to pain behaviour?)

**Dispositions**

As already implied, reference just to actual behaviour would not be enough for a plausible behaviourist analysis. I might not pull my hand away from the burning kettle despite desiring to I even more want to show my stoicism (or, of course, if it is tied to the kettle!). But I would have done so if it had not been fixed there or if my other desires had been different. Such conditionals ‘If circumstances have been thus and so, then I would have done such and such’ report dispositions I have. When I desire to move my hand, I am disposed to act in a certain way if the circumstances are right. Having this disposition is what it is for me to have the desire.

For Ryle, having a disposition is just a matter of certain conditional sentences being true. Cf. something’s being soluble is a matter of what would happen to it if it were put into water: i.e. it would dissolve. But this isn’t the only, or best, way of thinking of dispositions: on a realist view dispositions are truth-makers for the relevant conditional sentences: for X to be soluble is for X to be such that if put into water, X will dissolve. Consider ‘it dissolved because it was soluble’: on the realist view, this is ‘it dissolved because it was such that if put in water it would dissolve’—leaving room for further description of the causally relevant state gestured towards by ‘such that’. On Ryle’s view, by contrast, the claim is just ‘it dissolved because that is what it does when put in water’, which is not a causal claim at all. If beliefs are dispositions, best to think of them as dispositions realistically construed, else beliefs won’t be causal.

Mental states, if behavioural dispositions, are multi-track dispositions. Believing that the ice is thin is being disposed not to skate if one wants to stay dry, to skate if one wants to
drown, and to shout a warning if one wants to save Fred, and to keep quiet if one wants him to drown, and so on and so forth.

From Behaviourism to Functionalism

The reductive behaviourist’s ambition is to translate away claims about beliefs (desires, etc.) into long iffy claims about behaviour. But this can’t be done: for any prospective translations would require reference to other mental states. Whether I would go to the fridge if I believe there to be beer in it depends on whether I desire beer and whether I know where the fridge is and whether I believe I won’t be stopped on the way there.... So we must abandon this sort of reductive ambition.

Still, that doesn’t stop us identifying beliefs as dispositions, realistically construed as inner states identified by their role in producing behaviour and other mental states – the functionalist view. The conditionals characterizing the having of certain beliefs will specify behaviour given certain other mental states: the general principles used to spell out the right conditionals are principles of Folk Psychology.

And what distinguishes e.g. beliefs from desires? Beliefs might be distinguished from desires through their typical causes, in particular beliefs often have perceptual causes. Desires on the other hand are often caused by deprivation (e.g. desire for food). Hume thought only desires (and not beliefs) were motivational—i.e. actually motivate the subject to act.

NB: functionalism (the view that what makes something a belief is its functional role) is not to be confused with hierarchical functionalism (a view about how various enquiries into the mind/brain fit together).

Reading

Introductory
S & J: Chapter 10 and 11.
Churchland MC: Chapter 2.
Davies PM: §§1.2–1.5.

Key reading

Hume on Belief
D. Hume: A Treatise of Human Nature Book I, Part III, Section VII; and Appendix to the Treatise.

Behaviourism

J.B. Watson: An excerpt from ‘Talking and Thinking’ in Lycan MC.
R. Carnap: An excerpt from Psychology in Physical Language; in Lycan MC.

6. Levels of Description

Let’s sketch what might (very pretentiously!) be called ‘The Grand Unified Programme of Cognitive Neuroscience’ (‘neuroscience’ is, of course, the general scientific study of the brain: ‘cognitive’ signals that we are particularly interested in the investigation of how the brain processes information in perception, memory, and so on). This Grand Programme provides the best framework for explaining how different levels of description of the mind and brain hang together. For reasons that will soon become clear, the Programme can aptly be called a ‘functionalist’ one. We’ll take examples to do with consciousness.

Transitive vs. Intransitive Consciousness

We can usefully distinguish ‘transitive’ from ‘intransitive’ uses of the verb to be conscious. Those labels are grammatically a bit inaccurate: but they indicate the distinction between cases where we say that someone or something is conscious, full stop, and cases where we say that someone is conscious of something or conscious that something is the case. Let’s begin by considering very briefly the intransitive case.

In its familiar and central sense, to be conscious is just a matter of being awake, alert, non-comatose: thus, we may ask whether Jill has come round from the anaesthetic by asking ‘Is she conscious yet?’ A related intransitive use is in play when we ask, in a more philosophical tone, whether dolphins or mice or bluebottles are really conscious: do they actually have an inner life of the kind that we enjoy when we are awake and alert? Now, it is tempting to suppose that consciousness in the central intransitive sense is an all-or-nothing business; either the inner light is on (so to speak) or everything inside is dark. But this supposition can hardly survive much serious reflection. For consider: as we descend the phylogenetic tree, the mental capacities of animals become less developed and less complex— but there is no point on the downward path where we can plausibly draw a line and say ‘Above this point, animals have an inner life bathed in the bright light of consciousness, and below this line all is dark.’ Again, consider the rather depressing future that perhaps awaits some of us—severe mental decline due to neural degeneration. It is hardly plausible to suppose that, as the mental faculties decline and we sink into a final torpor (become, as they say, just a vegetable), there is a decisive moment when the bright light of consciousness is suddenly and permanently extinguished in us. In fact, the very idea that we can draw a sharp line here is no more plausible than the old Cartesian idea that Tibetans have an immaterial soul and tomatoes don’t, and somewhere between the two there is a sudden cross-over between organisms that do and organisms that don’t have a soul.

Of course, it is very difficult to imagine (imagine ‘from the inside’) what it is like to suffer from seriously impaired mental functioning. It is very difficult, perhaps impossible, to imagine the fragmentary mental world of a patient suffering from the late stages of Alzheimer’s disease. There is a principled reason why this is so: putting it crudely, imagining ‘from the inside’ involves the recapitulation and reorganization of experience—and hence the range of our imagination is limited by the raw material provided in our own mental life. We can perhaps imagine forms of consciousness slightly different from our own; but otherwise we tend to be imaginatively all at sea, and draw a blank. Hence, as far as our imaginations can take us, the alternatives seem to be either a full conscious life much like our own or no conscious life at all; no wonder that we so easily suppose that consciousness is a simple property which a creature either definitely has or definitely lacks. But that is assuredly a mistake. The term ‘conscious’ in fact roughly indicates a syndrome of mental conditions, conditions which can be present to various degrees, and which can be peel apart in various ways. Consider, for example, receptiveness to various kinds of perceptual stimuli, the functioning of the processing mechanisms which enables some sense to be made of these stimuli, the readiness to engage in discursive thought, the formation of reflective action-plans—these constituents of conscious life can evidently be present to different degrees of development as we descend the phylogenetic tree. And while these constituents do normally go together for rather obvious reasons (there is, for example, little evolutionary point in being able to think if you have no perceptual input to process or if you never form intentions to act) they can begin to come apart in cases of sufficiently severe neural damage or degeneration.

Normal human consciousness, then, is a complex structured state, not a simple unitary one. It standardly involves some awareness of one’s environment—and it involves too some
awareness of one’s own thoughts and sensations. Putting it more formally, the intransitive notion of consciousness plausibly invites analysis in terms of (among other things) transitive consciousness—that is, in terms of the notion of being conscious of something. So let’s turn next to say something equally brief about some of the transitive cases.

Perception

Consider first the case of perceptual consciousness of one’s environment. Once upon a time, it might have been possible to think of visual perception (for example) as a relatively straightforward matter, involving the triggering of a play of uninterpreted images before the mind’s eye. But those happy days are gone; the psychologists have long since convinced us of the complex cognitive structure of vision. And this complexity is dramatically confirmed in the neurology clinic, where we find patients with selective cognitive deficits caused by various kinds of localized neural damage. Take visual agnosia, e.g. the case described by Oliver Sacks’s title essay in his book The Man who Mistook his Wife for a Hat. A more detailed and experimentally controlled investigation into a patient with visual agnosia has recently been reported by Glyn Humphreys and Jane Riddoch. Their patient has normal acuity of vision, and can accurately make very detailed copies of pictures, line by line. He can see bit by bit the fine detail of a face (for example). He can also interpret silhouette outlines without problems, showing that he has a good grasp of visual information about global form. But he cannot integrate his visual information about small features with his visual information about global form—so (like Sacks’s patient) he cannot recognize, for example, faces or places. If he looks at his own face in the mirror, he sees it either as sets of unrelated details (hairline, scar on the nose, or whatever) or as a schematic or ‘global’ face; what he cannot do is what the normally sighted do, namely see a face as a whole, but with the details elaborated.

To take another quick example from the neurology clinic: Zihl, von Cramon and Mai describe a patient with a very different visual disorder—an inability to see movement (other than the slow movement of objects bang centre in the field of vision). Again, basic visual acuity is good: but the patient has difficulty in pouring tea into a cup, for example, because the fluid appears to be frozen, like a glacier; and she cannot stop pouring at the right time since she is unable to perceive the movement in the cup when the fluid rises. She finds it uncomfortable to be in a room where others are walking about, because people suddenly pop up here or there although she has not seen them moving. Understandably enough, crossing the road is a nightmare.

Such cases show us something of the structure of our visual processing they reveal that the processing of global form information, the processing of the detailed articulation of a form, and the recognition of movement are initially segregated into different cortical areas. Neural damage can knock out one of these processes, or stop informational output from one process being available as input to some higher level process. The discovery of this sort of structuring is very illuminating, though it can initially seem highly counter-intuitive. People are sometimes inclined to protest ‘If someone can see detail perfectly well and can see outlines perfectly well, then (dammit!) how can he possibly not be able to make out a face? Likewise, if someone can see a scene, how can she possibly miss the movement (unless she is intermittently blind)? These so-called visual deficits seem to be visual nonsenses.’ And there is a very real point to such protests. Our vernacular concept of seeing has its home in describing the normal, everyday cases: if we try to apply it to the abnormal cases, we can indeed find ourselves of tripping over own linguistic bootstraps. The moral is—don’t expect to learn too much from reflecting on ordinary language.

Vision, like the other modalities of perceptual consciousness, is much more complex than we might perhaps have supposed. How, then, are we to approach its study? Well, that is a pretty presumptuous question: after all, the neurophysiologists, the neurologists, the psychologists, the cognitive scientists, the artificial intelligence buffs, are already beavering away investigating the visual system—and who are philosophers to tell them how to do their jobs? So rephrase the question: how are we to integrate these various studies of the visual system into an overall picture?

Hierarchical functionalism
A plausible answer starts something like this. The visual system as a whole has certain overall functions—most fundamentally, that of recovering detailed information about the three-dimensional world from the two-dimensional arrays of stimuli to the retinal surfaces. We want to understand how the trick is done; and the natural way to proceed is to conceive of the overall system as comprising a number of sub-systems which perform more limited functions (as it perhaps might be, a ‘first pass’ global form detecting system, a ‘second pass’ system which elaborates global form information with codings of local features, and so on). But now of course, we want to understand how these more limited functions are performed; and again the way to proceed is to see these functions as performed (as it were) by a co-operating team of somewhat less clever functionaries. The global form detecting system might perhaps involve edge detecting systems, vertex counters, a system that matches edges and vertices with stored shape templates, and so on. And in its turn, to go down one more level, an edge detector may comprise—among other things—arrays of units whose individual function is to respond differentially to light/dark boundaries which have a particular orientation. And here our story nicely links up with neurophysiology: for we know that there are indeed cells in the striate cortex which have just this sort of role, of being excited by (say) an oblique light/dark boundary falling on the retina.

There is, of course, nothing mysterious about the idea that we can gain an understanding of the workings of a complex system by seeing it as built up simpler functional units, which are in turn built of yet simpler units. A child comes to you wanting to know how a car functions: you start by splitting the car into a number of large-scale functional units, distinguishing the engine, the steering, the braking system, and so forth. ‘Ah, but how does the engine work?’ You respond by going down to a more detailed level of functional description, and subdividing the engine system into a fuel supply, fuel mixer, a firing system, etc. ‘Yes, but how does the firing system work?’ Well, on you go, down to a yet lower functional level, and you speak of condensers, distributors, sparking plugs, and the rest. And so it goes. We come to understand how the car works by working our way around the hierarchy (the pyramid, if you like) of functional levels, and come to see at each level how various functions are performed by systems of still simpler units.

Similarly with the visual system. This too is a complex and exceedingly clever functional system (though a naturally evolved system rather than an intentionally built one like a car): to understand how it performs, we need to see it as comprising sub-systems executing somewhat less clever tasks, sub-systems which are in turn comprised of simpler functional units, and so on downwards, until we reach the level of very simple functions which are not clever at all, and which can be performed by relatively simple devices like smallish nets of neurones. The neurophysiologists and their other scientific colleagues can then be regarded as exploring various levels of this hierarchy of functional structures. Cognitive scientists and artificial intelligence workers on computer vision are working near the top level, investigating the sorts of high-level functional structures that could perform the tasks of the visual system; neuroscientists are mostly working from the bottom up, telling us about neurones and small neural nets; investigators into the parallel distributed processing potentialities of larger neural nets are working a couple of levels up from the bottom; and so on. So we have a Grand Unified Programme for research in vision, with different workers digging away at different levels of the functional hierarchy.

(NB: a warning. Don’t confuse hierarchical functionalism as a story about how different enquiries into e.g. vision fit together, with functionalism as a story about the analysis or ontology of folk-psychology. More about this later.)

Self-awareness

What goes for vision goes too for other kinds of mental transaction: a hierarchical functionalist programme looks inviting. Consider in particular our consciousness of our own minds, the introspective awareness of what we ourselves are thinking and feeling (the richness of our self-awareness is perhaps what distinguishes us from lower animals). Traditional doctrine regards this as a pretty straightforward business. John Locke, for example, famously defined consciousness as ‘the perception of what passes in a man’s own mind’, and he adds that ‘it is impossible for anyone to perceive without perceiving that he does perceive’. Thus for Locke our self-awareness is a clear inner light which consistently reveals all. But this is pretty evidently untrue; much goes on mentally of which we are not aware. Recall the familiar and
mildly alarming experience of driving home engrossed in a conversation, only to realize with a start that you have just driven a mile or two along a twisting road without noticing what you are doing: you must have perceived the road (or you would have run off it long before), but you certainly didn’t seem to perceive that you were perceiving the road. ‘Still,’ it might be said, ‘even if not all one’s experiences clamour for conscious attention, it remains true that when you do turn your inner gaze on some aspect of your mental life, then all is revealed.’ But that isn’t true either. And it isn’t only the seamy side of our inner life which we hide away in Freudian dark corners: in fact a great deal of mental activity goes on unavailable to introspection. A rather nice and very simple example from the psychology laboratory concerns the role of our perception of pupil size in judging the friendliness of a face. Of two otherwise similar pictures, the one whose eyes have the greater pupil size is consistently judged to be the more appealing and friendly; but experimental subjects are not only unaware of using perceptual information about pupil size in forming a judgement, they are not aware of having perceived the difference in size at all. Introspection just doesn’t seem to reveal what is going on here.

The limitations and fallibility of introspective consciousness is confirmed more dramatically in the neurology clinic. Larry Weiskrantz has famously investigated the phenomenon of blind-sight; here brain-damaged patients who have no introspectible visual sensations, who report blindness, are found (unbeknownst to themselves) to have a residual visual discriminatory capacity. For example, they sincerely aver not to be able to see a flashing light; but when asked to guess in which direction the light is, they are able to give the right answer with impressive frequency. This phenomenon illustrates how we may acquire perceptual information although introspectively there seems to be nothing going on. And even odder is the converse syndrome of blindness denial. In the blindsight cases there is residual vision but no belief that one can see; in the case of patients suffering from Anton’s syndrome, there is the belief that one can still see but no vision. Patients who suffer a sudden loss of vision due to insult to the visual cortex will sometimes (with every appearance of sincerity) deny that they are blind, and systematically confabulate to explain away why they keep bumping into the furniture. If and when vision is restored, the patients persist that in their story that nothing has been amiss. Now, people may again be inclined to protest: ‘If someone can’t see, then—dash it all!—he must know he can’t’. And again the protest has a point: our common-sense conception of the way we know our own minds is indeed put under strain by such cases. But the moral is as before; our common-sense conception is not the last word. Our internal self-monitorings and self-understandings are much more fragmentary, much less informative, much less reliable than we like to think.

If Lockian simplicities are out, how should we understand our self-awareness? This isn’t to ask ‘What account can we give of self-awareness at the level of common-sense talk about persons?’ We can normally just say what we have been seeing (for example), when we are not otherwise distracted. This ability is, so to speak, a ‘given’ – not amenable to further analysis at the personal, common-sense level. But we can inquire about the sub-personal structures in virtue of which we have the ability. Perhaps it is like this: some of the output from our visual system and from our other perceptual modalities is passed to a short-term memory buffer that our central control systems can interrogate. There is competition for space in this buffer. Normally, when we are driving along, the visual information is given priority; but if our conversation is particularly engrossing then it is the auditory information that claims space in the buffer, and therefore we cannot introspectively recover the excluded perceptual information that our motor co-ordinating system used for governing our movement of the steering wheel. Again, in the blindsight patients studied by Weiskrantz, their residual visual processing does not deliver recoverable output to the buffer.

It is doubtless a very great deal more complicated than that. Still, this surely is the beginnings of the sort of story we want (if we are to proceed beyond the blankly unhelpful deliverances of the common-sense description of the phenomena)—a story about the interplay between various high level processing systems realized in the brain. We want to know about the sorts of output from different level processes that can be made available for recall by our central processing units, we want to know the kinds of filtering and censoring functions that can be in play, and of course much more besides. In short, what we want is a nice rich hierarchical functionalist model of our introspective capacities, to augment our functionalist story about our visual system and the rest.

Minds and brains
In bald summary: Consciousness (intransitive sense) is a syndrome of interconnected phenomena, including in particular the varieties of transitive consciousness. And these different ways of being conscious (transitive sense)—whether being conscious of the external world or being self-conscious—are in turn very complex, to be understood by articulating accounts of how various cognitive and perceptual sub-processes are hierarchically organized to interact with each other. From this perspective, consciousness is a problem all right. No one says that it is going to be easy to articulate—even in rough outline the desired accounts of our mental organization, or to discover how that organization is realized in the brain. We may have a Grand Programme for co-operative research in cognitive neuroscience; but no one supposes that following through the programme is simple. Still, we have promising research strategies, and there is no reason to despair.

However, all this is likely this to meet with very vigorous protests. ‘It’s no doubt a fine and splendid thing,’ the argument goes, ‘to try to develop an account of the functional structure of the brain; but it won’t tell us about the mind.’ Yet why so? The full story about the brain will tell us about nothing other than the processes and states that we ignorantly gesture at with our everyday psychological vocabulary. We talk about beliefs and desires; the top level of the hierarchical story deals in information-registering states and goal orientations. We talk of pains; and a rather lower level in the story gives a functional characterization of the neural workings of our damage-monitoring systems. This isn’t to say that we can line up our folk talk against the scientific story, and find absolutely neat correlations. The match between folk talk and reality is no doubt a bit messy and complex; but then the everyday purposes of folk talk are themselves messy and complex, so there is no surprise in that. And this is no reason for trying to get on without common-sense psychology: it is no doubt splendid in its place but its place isn’t telling us systematically and in explanatory detail what makes us tick. Folk talk only gestures towards what work on the Grand Unified Programme can deliver, a systematic account of the mind.

Reading

For the basic picture of hierarchical functionalism, see

For some of the neurological information alluded to, see the fascinating
7. Minds and Machines

Comparing Minds and Machines

Hilary Putnam once defended a comparison between mental states and the states of Turing machines (an idealized digital computer).

A Turing machine has various internal states, takes inputs and produces outputs. The output produced and the next state the machine goes into are determined by the input and the current internal state of the machine. Instructions governing the machine can be stated in the form: ‘if the input is I and the current state $S_i$, then output $O$ and go into state $S_j$.’ And there will be an instruction of this form for each pair of input and current state. A Turing machine is characterized by its list of instructions. We can display them in the form of a machine table with columns for each state and rows for each input, with each square containing the relevant pair of output and new state. The essence of a state of a Turing machine is its role as characterized by that table. What realizes the internal states in any concrete realisation of a Turing machine is irrelevant. And machines that are internally very different can have the same machine table and the same functionally characterized states. (Familiarly, computers with different chips can run the same C program.)

Putnam’s suggestion is that mental states are like the states of a Turing machine. What is distinctive about a mental state is its role in the production of behaviour (outputs) given sensory inputs. I am in pain when I am in a state which plays a certain role as specified by a certain machine table. What physical (or nonphysical) state actually plays the role is irrelevant to the ascription of pain.

Putnam’s position is often referred to as Machine Functionalism. The immediate analogy is that: the mind is to the brain as a computer programme is to the hardware of the computer on which it runs. States of the programme (identified by their role in its execution) can be specified by a Turing machine table and the physical state playing that role will be irrelevant. Similarly, the subject-matter of psychology is the functionally identified psychological states which realize the ‘psychological programme’. Psychology need not be concerned with the physical state which plays that role. Psychology explains behaviour by reference to sensory inputs and functionally identified mental states. Characteristic explanations in psychology need to work at this higher level of abstraction, and could not be given with reference to neuroscience of the brain alone.

Putnam presented his position as opposing all identity theories. Is this right? Whether functionalism conflicts with materialism could depend on what is understood by materialism. If the emphasis is on whether the only things which exist are material, the positions seem compatible. If materialism also requires that all properties are physical then arguably functionalism must reject it: a given functional property cannot simply be identified with any particular physical property.

How close is the analogy?

Can mental states actually be identified with machine-table states for the ‘programme’ which we are ‘running’?

1) The same bit of hardware can be regarded as (simultaneously) realizing many different Turing Machines. Take the minimal machine, with just one input state $I$, one output state $O$, one inner state $S$, and the rule “given $I$, output $O$, and stay in state $S$”. Then we all realize this machine (just take any input as counting as realizing $I$, and any behaviour as counting as realizing $O$!). So of the abstract machine-tables we exemplify, whose states correspond to our mental states?

2) A Turing Machine (identified abstractly) is only in one state at a time: but we are in many mental states (in pain, thinking about Venice, in love with X ...). So the identification would have to be of total mental state with a Turing Machine. But then, in this picture, what corresponds to the thought that our total mental state has many parts, some of which recur form time to time? (I’m thinking about Venice on many different occasions.) It is crucial to belief-desire psychology to think of our total mental state as factoring into substates; the Turing machine analogy doesn’t capture this.
Still, we might think that idea of computational states is useful analogy: the software/hardware model helps us understand multiple realizability, why mental explanations are illuminating even if we know the hardware details, etc. (Also, incidentally, it gives nice model of introspection. The machine which has the program rule ‘When in S, print “I am in S”’, will have a sort of privileged access to when it is in S—it can just say that it is, without having to engage in any processing.)

Troubles for the mental/computational model?

However, even if we retreat to a looser comparison between computational and mental states, there are problems. The notion of a computational state is too abstract, it will be said; so computational functionalism would allow too much to count as a thinker:

Block’s Chinese Giant: Suppose we get the population of China to act out the behaviour of a set of neurones, running the right ‘program’ (or what about a set of beer cans ingeniously rigged up with string so that disturbing one, the effect ripples across the collection in the programmed pattern, or ...?), and you’d have a thinker!

Searle’s Chinese Room: Get someone to hand simulate a program—e.g. the ‘speaking Chinese’ program—and you’ll have a Chinese speaker. You post squiggles through a letter box: I dash madly around, following a program that tells me, in the end, if I receive “squiggle, squiggle”, to respond “squoggle, squoggle”. So I post back squoggles. Unknown to me, I am apparently responding to questions in Chinese (passing the “Turing Test” for intelligence).

1) To Block: But don’t get distracted by size-ism. If you were miniaturized, and found yourself wandering in my brain, it would seem odd to reflect that these neurones you are encountering were responsible for thought. Suppose my neurones are dying off, and one by one they are replaced by more miniaturized persons, working artificial neurones according to my current ‘program’. Wouldn’t I keep going as a thinker composed of some smaller thinkers? What makes us think that the Chinese Giant isn’t a thinker is that it isn’t correctly embedded in its environment in the way required for representational states — cf. Lecture Topic 2.

2) To Searle: It is people, whole systems, that think, speak Chinese or whatever. Look inside the room and there is no Chinese speaker (but there isn’t one in the brain either). Doesn’t follow that the whole system doesn’t speak Chinese.

Reading

abbreviations

On Putnam

Hilary Putnam: ‘Philosophy and our Mental Life’. In MLR, RPP.

On Block


On Searle


D. Hofstadter: ‘Reflections [on Searle]’, in Hofstadter & Dennett.


8. Functionalism

Unlike the Rylean behaviourist, the functionalist claims that a mental state is an internal state which causes behaviour. But it preserves the connection with behaviour by characterizing the state in terms of the behaviour which it causes. Functionalism also allows reference to other mental states in the characterization of a particular mental state. A reductive definition in terms of observable circumstances is not expected. That the subject must have other particular beliefs and desires if a given state is to cause a given action, is part of what defines that belief.

The role of a state will be given by:

1) Its typical causes (e.g. the fact that damage to body-tissues is a typical cause of pain).
2) The ways it combines with other mental states to produce behaviour.
3) The ways it combines with other mental states to produce other mental states.

So, for example, believing that it is raining is being in that state which is typically caused by rain falling in the proximity of the subject; and which when combined with the desire not to get wet (and certain other beliefs) causes the subject takes an umbrella; and which when combined with certain other desires and beliefs causes the subject to bring in washing; and ... (A long list of such generalisations is required to identify each mental state.)

We might similarly essay a functionalist account for non-intentional states. Pain is that state which is typically caused by damage to the skin; and which typically causes raised blood pressure, and wincing; and which combines with other certain beliefs to cause the subject to visit the doctor; and which causes the desire for the pain to stop... (See S&J, Chapters XIV and XV – and Special Topic A for more on this.)

So mental states are internal states of the subject which play a causal role between sensory input and behavioural output. Each state is identified by its unique role it plays in this mediation: the distinctive ways in which it combines with other states to produce given outputs. This is referred to as its functional role.

Functionalism is compatible with materialism or with dualism (assuming that dualism is coherent, or it isn’t compatible with anything!) For a mental state is whatever has the specified causes and effects. That could be a particular state of the brain, or it could (maybe) be a property of an immaterial mind. But functionalists are most commonly materialists. If the internal causes of our behaviour are brain states, then there is no need to postulate Cartesian mind-stuff to account for our mental lives.

Functionalism is often described as having its roots in Aristotle’s philosophy. He draws a distinction between the matter and the form of a thing. The form of something is what makes it what it is. For example, the form of a knife would be its capacity to cut: something is a knife in virtue of having this capacity. The matter of something is what it is made from. So the matter of the same knife could be a piece of stainless steel.

Aristotle took the mind or soul to be the form of a human and explained that this was its capacities for the activities which are characteristic of rational humans (the matter of humans is skin and bones etc.) That is, what it is to have a mind is to have certain capacities. (And these are not entities). The mind itself is not a thing. Rather, having a mind is having a certain range of abilities. [Compare an account of what it is for something to be alive which draws on its capacity to work under its own steam.]

This line is naturally developed – and is by Aristotle – in a broadly functionalist direction: he takes what is distinctive about the mental to be its ability to bring about certain things, in particular behaviour. What is distinctive about a given state is its role in that production of behaviour—its functional role.

See S&J Chapters XII–XIII for basic discussion. These notes discuss further the relation between functionalism and identity theories.

Combining an Identity Theory with Functionalism.

According to functionalism, what is distinctive about a mental state is its functional role. If it is always some state of the brain that plays that role, then the thesis that each token mental event is identical to a token brain event (token-token identity) should be upheld. Type-type
identity will follow if it is always the same type of brain state which plays a given functional role. Functionalism would answer questions of which brain state is identical to a given mental state by identifying which brain state plays the relevant functional role.

To illustrate a functionalist token-token identity theory: David’s belief that broccoli kills is a token of the state of believing-that-broccoli-kills and could be identical to a particular state of David’s brain at the time of the belief. Similarly Hilary’s belief that broccoli kills is a state of Hilary’s brain. But the brain state in the two cases may not be the same. What Hilary and David have in common in virtue of which they share a belief is not a type of brain state but a functional state.

Token-token identity theories do not face the problem of multiple realization faced by type-type identity theories since they do not deny that the same mental type can be realized by tokens of different physical types. But perhaps they overshoot: after all, we do research on the biological bases of pain (which expect anaesthetics to work on anyone: i.e. we expect some neural similarities in the brain-states related to pain). Again, we do research on the brain location of different memory functions, etc. So perhaps we should adopt the position that:

\[ \text{a mental type for a species} = \text{a physical type for a species} \]

Members of the same species which share a mental state share a physical state. But different physical states can realize the same mental state in different species. It is a narrower mental type that is identified with a physical type. E.g. David Lewis holds that what picks out the same mental state across species is the functional definition of that state. Thus he is a functionalist as well as an identity theorist. Before considering why he moves to these narrower species-relative types, we can look at his argument for the Identity Theory.

DAVID LEWIS’s Argument for a Restricted Type-Identity Theory.

First a general argument for an identity theory of one sort or another:

First Premiss: ‘The definitive characteristic of any experience [or other mental state] as such is its causal role’. A functionalist premiss. It is neutral between certain theories of the ontology of mind, but rather is offered as an account of what we mean by our expressions for mental states. Mental states are whatever plays the right causal role. These defining features allows the reality and causal efficacy of mental states and allows them to be interdefined [contrast Behaviourism]. Furthermore, they are defined by their typical causal roles, which may be absent on odd occasions (for the state will have other non-defining characteristics which will account for possible atypical causes or effects) [furnishing a response to the paralytic/perfect actor objections to Behaviourism].

Second Premiss: ‘There is some unified body of scientific theories, of the sort we now accept, which together provide a true and exhaustive account of all physical phenomena”. [Compare the principle of the closure of physics wielded against interaction between matter and Cartesian minds.] This denies that we need ever explain physical phenomena by nonphysical ones.

Aspects of behaviour are among the typical effects which define mental states. Since behaviour is a physical phenomenon it has purely physical causes (by the second premiss). So those mental states causing it cannot be non-physical. Hence they are identical to physical phenomena of some sort: in particular they will be neural states. Hence Conclusion: Mental states are some physical phenomena or other.

But this leaves it open whether to go for a universal type theory, a restricted, species-relative type theory, or a token-token theory.

Why Species-relative types? Mad Pain and Martian Pain

Lewis considers two sorts of cases in which we would attribute pain to the subject but which a pure type-type or pure token-token identity theorist may be unable to accommodate.
The Madman. Lewis claims that there could be a ‘madman’ who feels pain, but whose pain is not caused in the normal way and who does not respond to it in the normal way. E.g. this madman’s pain is caused by moderate exercise on an empty stomach and causes him to cross his legs and snap his fingers. His pain does not occupy the usual causal role, but, Lewis claims, it could still be pain he feels in these situations. A functionalist token-token identity theory would have to deny that he feels pain in these circumstances since his state does not occupy the right functional role. By adopting a form of type identity theory, he can explain that the madman is in a state of pain because he is in the neural state which ‘normal’ members of his species are in when in pain.

The Martian. Lewis claims that a (normal) Martian would be in pain if in the same functional state as a (normal) human. But that Martian would not be in the same physical state (assuming its brain is nothing like ours). A type-identity theory which was not species relative would have to deny that the Martian was in pain. Hence we need to go species-relative.

NB Mad Martians can also be accommodated by inheriting their mental states from states of normal Martians in the same physical states. But Lewis has to deny the possibility of a species with only one member who is mad. Lewis does not consider this to be a very damaging objection.

DONALD DAVIDSON’s argument for a token-identity theory

Davidson’s theory of mind is a token-token identity theory. We will look at one argument he offers for his position (in his ‘Mental Events’) and at more general issues raised by his discussions.

The premisses:
1) At least some mental events causally interact with physical events.
2) Where there is a particular relation of cause and effect it falls under a general law of nature.
3) There are no strict laws on the basis of which mental events can be predicted and explained.

The apparent inconsistency is resolved by concluding:
4) (Token) mental events are identical to (token) physical events.

(1) to (3) threaten to be inconsistent by the following argument: take some particular causal interaction whereby a physical event, P causes a mental event, M—(1) guarantees that such interactions exist. This interaction must fall under a general causal law, by (2). This law could then apparently be used to relate P-type events to M-type events. But (3) says that there are no such laws. (4) can resolve the apparent contradiction in the following way: if every mental event is identical to some physical event or other then P can cause M in virtue of a law which picks out some physical aspects of the event M; but different instances of M will be tokened in different ways, and different laws will be in play on different occasions of M-causation. Davidson denies that there could be reduction of the mental to the physical, as demanded by type-type identity theories. Such a reduction would require laws relating the mental and the physical, e.g. of the form ‘if x is P then it is M’ (bridging laws).

Davidson gives his token-identity theory the fancy title of Anomalous Monism. It is called monism because it is only committed to one kind of substance (matter) in contrast with dualism’s two kinds. The ‘anomalous’ indicates his thesis that there are no mental laws; here anomalous = lawless.

(1) is hardly deniable – and is, of course, a central functionalist claim. (2) is a widely accepted general principle (if not entirely beyond dispute). Davidson does not deny that you can know that one particular event causes another without knowing the general law instantiated, but, he claims, there must be such a law for every causal relation. The principle requires that if event c causes event e, then for some description of c and some description of e, it is a law that anything fulfilling the first description causes something fulfilling the latter description. Not all descriptions of c and e will figure in a law statement, the principle only requires there is at least one pair that does. (Perhaps the event described on page 3 of the Times caused the event described on page 10: but there won’t be laws relating page-three events to page-ten
events, at least not as so described.) Davidson holds that it is a physical description of a mental event that figures in the law statement when that event causally interacts.

(3) is the really controversial principle, then. Davidson considers science to be essentially the study of laws, so in asserting (3) he is committing himself to the claim that psychology cannot be a science. So the central question is ...

Could there be Psychological and Psychophysical laws?

(Psychological laws would involve just mental – psychological – events, whilst psychophysical laws would relate the mental and physical.)

NB Law statements entail true, unrestricted empirical generalizations. So if "all ravens are black" states a law then it must be true of every raven that it is black: the generalization must be without exceptions.

They must also support counterfactuals. If it is a law that all As are B then it is true of anything that is not an A that if it had been an A it would also have been B. (Thus ruling out as law statements accidentally true generalizations such as ‘all people in this room speak English’.)

Note that the possibility of laws involving the mental is prima facie plausible. E.g. the operation of anaesthetics assumes that physical events of a particular type will cause the right mental event (e.g. render the patient unconscious).

Davidson allows that there are generalizations involving the mental, just that they couldn’t be lawlike. In particular, he argues that many generalizations involving the mental will be only roughly true (not exceptionless, as genuine laws would be). True, we can gerrymander some horrible complex accidental generalizations that are exceptionless. In particular, if there are only a finite number of events falling under a given mental predicate, then by listing physical properties of those events we could, in theory, give a true generalization about physical properties of events falling under that mental predicate: but it still won’t be lawlike, as it won’t sustain counterfactuals.

Why does Davidson think that generalizations involving the mental will (almost) always be only roughly true?

a) Something could come in the way to prevent a mental event having its expected effect. E.g. other mental states could prevent me executing my decision: no law can govern it.

But similarly for many other laws. E.g. though it is a geological law that a meandering river erodes the bank, this can be obstructed by other facts. It is even the case that something will not fall to earth if another force diverts it, despite the laws of gravity.

If there were mental laws they would need a qualification ‘other things being equal’ (a ceteris paribus clause). But this is a common feature of non-mental laws.

b) But if we were to spell out that ceteris paribus clause in the mental case we would have to refer to the non-mental (for the effect could be prevented by a physical obstruction). So there can be no purely mental laws, mental generalizations are ‘heteronomic’.

But similarly chemical, biological or geological generalizations (those from other ‘special sciences’) would need reference to physical facts. Are we to believe that there are no laws in these realms either and that these are not sciences either? Moreover, even laws of gravity might count as ‘heteronomic’ by the same argument, since the laws only apply to ensure an acceleration of X cm/sec/sec so long as e.g. there is no interference by electrical forces, for example.

c) Other reasons Davidson has for denying that there could be laws involving the mental draw on distinctive features of the mental, in particular the role of rationality and holism. He is particularly concerned with those mental states with content: beliefs and desires etc.. He argues that the mental states we attribute to someone are constrained by coherence and rationality assumptions. They must be part of a coherent, rational system of mental states and cannot be considered in isolation from the whole of the system (holism). If it is an essential feature of the mental is the link to rationality, but this is not a feature of the physical, then
laws relating the two may be unable to accommodate this discrepancy. (But why so? And NB, this argument poses no threat to laws involving mental states without content, such as pains. We could agree with Lewis and take a restricted type-theory of those.)

d) Davidson further believes that there is indeterminacy of interpretation: there is no unique correct interpretation of someone’s beliefs and desires. If beliefs do not have determinate content then how can there be cannot be laws about all beliefs with a certain content?

Does Davidson really respect the efficacy of the mental?

One objection which has been raised against Davidson runs as follows: He cannot accommodate the fact that it is ‘the mental as mental’ that has causes and effects in the standard examples of mental-physical causation. My decision to go skiing causes me to go skiing in virtue of the fact that the former event was such a decision, i.e. in virtue of that mental property of the event. Davidson, on the other hand, must say that the causal relation is due to a physical property of that event: it is ‘the mental as physical’ which engages with the laws and which has effects. Hence, the objection goes, for Davidson the mental properties of events are causally irrelevant and there is only an accidental connection between the mental and the physical. His position seems like a form of epiphenomenalism (or worse, since the mental as mental cannot have causes either):

Adopting the functionalist strategy of identifying a mental event as the state with certain causes and effects might dissolve this objection. For then it would not be merely an accident that the event with those mental properties has the effects that it does. For which (token) physical event is identified with the token mental effect is partly specified by the fact that it stands in the relevant causal relations. NB this functionalist identification of mental states does not commit you to psychophysical laws.

Supervenience of the Mental on the Physical

To respect the connection between the physical and the mental, allowing the mental to be in some sense dependent on the physical, Davidson holds a supervenience thesis. This is weaker than the demand for laws relating the two, and does not imply that the mental is reducible to the physical. [The thesis is fairly widely accepted among contemporary philosophers.]

Supervenience of the mental on the physical:

Two things cannot exactly resemble each other in all physical respects whilst differing in mental respects.

Hence once the physical properties of something are specified, its mental properties are fixed. (Cf. supervenience of moral facts on physical facts)

Many problems remain...

a) The supervenience claim is compatible with the possibility that two people could differ in physical respects only in the length of a single eyelash and yet have different mental states. This seems absurd. But physical properties on which the mental depends cannot be singled out without drawing on psychophysical laws.

b) Might the supervenience thesis also commit you to psychophysical laws after all? Suppose I have a set of physical properties, P. By supervenience, anyone who has those properties would also share my set of mental properties, M. And this generalisation has counterfactual support. Thus it seems to be a law that anyone with P has M. [NB this need not give reduction of M to P, for someone could share all my mental properties, whilst having different physical properties compatibly with supervenience.]

Reading

Abbreviations


Introductory

S & J: Chapters 6, 12 and 13.
Churchland MC: Chapter 2.3 and 2.4.

N. Block: ‘What is Functionalism?’ in RPP.

On Lewis

D. Lewis: ‘Mad Pain and Martian Pain’ in D. Lewis Philosophical Papers Vol 1. Oxford University Press, 1983 (including a postscript). Also in Block RPP.


On Davidson
D. Davidson: ‘Mental Events’ in NM, RPP. Also in Davidson’s Essays on Actions and Events, Clarendon Press, 1980. [O]
D. Davidson: ‘Philosophy as Psychology’, in his Essays on Actions and Events..
9. Realism, Instrumentalism, Eliminativism

1. Folk Psychology

As noted before, ‘Folk Psychology’ is a label for the body of our generally accepted, common-sense, principles which relate mental states to one another and to behaviour. Examples may include the generalization that (often) if someone is hungry, she is irritable, or that someone believing that it will rain and desiring not to get wet will e.g. avoid venturing out.

On the basis of such principles we can predict and explain actions. E.g. we predict how someone will act on the basis of knowing some of her beliefs; or we explain someone’s irritability by the fact that they are hungry. If people who believe that p tend to act in particular ways, then Arnold’s believing that p is a good explanation of why he acted that way; or knowing that he believes that p gives good reason for predicting that he will act that way.

The most general principles governing beliefs and desires will apply to a whole range of different states. E.g.:

(1) If someone desires that p and believes that doing A will bring about p, she will normally do q (provided she has no stronger conflicting desires etc.).

This enables me to predict that you will take your umbrella out (action A) if you desire not to get wet (desire p) and believe that taking your umbrella will stop you getting wet. Another example:

(2) If someone has the belief that p, and q follows from p, that person will normally tend to believe that q (assuming that the inference is fairly straightforward).

Other principles, such as that hunger causes irritability etc., are more specific. Taken together, these principles of Folk Psychology specify the functional roles by which the functionalist identifies mental states. Having a particular belief is being in a state which stands in those causal relations with actions and other mental states as specified by Folk Psychology.

We make enormous use of Folk Psychology in our everyday lives. True, we rarely cite a general principle like (1) or (2) when we do so, and tend not to be conscious of the principles at work. In general, it is implicit knowledge that people have of them. This can be compared with our knowledge of grammatical rules, which few of us could formulate despite implicit knowledge revealed by our ability to recognise, in general, whether or not a string of words satisfies those rules.

The principles of Folk Psychology are somewhat imprecise and contain hedging expressions such as ‘normally’. In certain unusual circumstances someone may not perform the actions those principles would lead us to predict: but this need not undermine the utility of those principles.

Prediction and explanation according to the principles of Folk Psychology, then, has impressively widespread success. On the whole, we tend not to be completely surprised by people’s actions (at least if we have some idea of ‘where they are coming from’ i.e. some idea of their beliefs and desires) and we tend to understand why they did what they did. The principles may not be sure-fire, but we can usefully rely on them.

Contrast this with trying to provide explanations and predictions of human behaviour by reference to their physical states only. This will be impossible in practice, because of the staggering complexity of the physical systems involved. Millions of neurons and other particles would be involved in any such a calculation of behaviour. And then there is the problem of predicting the future course of external stimuli.

2. Is Folk Psychology a Theory?

Folk psychology has various features shared by theories in general. In particular, it embodies a collection of general principles about the theoretical states postulated by the theory, i.e. our psychological states and activities, which are invoked to yield predictions and explanations. In making claims about the interconnections between various mental states and activities, our common-sense psychology is naturally taken as making claims about the real existence of these states and activities. For example, the principle stating “anyone with the belief that p will not normally have the belief that not-p” would surely be vacuous if no-one had beliefs. It might naturally be argued that in endorsing such principles, as folk psychology seems to, we
are committing ourselves to the real existence of states of beliefs and desires which obey the theoretical principles. Someone who thinks we should take these apparent commitments of folk psychology at face value might be called a **realist** about propositional attitudes.

But **should** we take these commitments seriously? If folk psychology is a theory, then—like other theories—it could be false. Genuine theories stick their necks out: their ambition makes them explanatory powerful, but also makes them vulnerable to falsification.

Now, we might wonder whether it so much as makes sense to suppose that our rough-and-ready folk principles about the mind could be discovered to be false. The idea that it makes sense to suppose that people like you and me don’t **really** have beliefs and desires can look bizarre. We are tempted to protest: “Look, there might be an issue about dolphins or Martians or computers: but there is no question but that you and I are believer/desirers. *That* isn’t a theory! For to be a believer/desirer just is to be something that goes in for the complicated kinds of interactions with the world that you and I go in for. Talk about beliefs and desires just is talk about these patterns in our interactions with the world.” The idea, then, is we should construe talk of beliefs and desires not as talk about theoretically postulated states, but as a compendious way of surveying patterns in behaviour. The ‘folk theory’ isn’t so much a theory (which might be true or false) as an instrument for seeing patterns in behaviour (which is acceptable so long as it works). On this **instrumentalist** view, the undisputed fact that folk psychology works well enough within its domain is reason enough to accept it.

What looks like a sophisticated version of this kind of instrumentalism has been defended by Daniel Dennett. In seeing something as a believer/desirer, we are ‘adopting the intentional stance’ to it; seeing it as behaving in a way that is rationally appropriate to its environment given its needs (and so, as a default, believing what it should believe about its environment, wanting what it should want given its needs). And if the adopting the stance works as applied to X, then that is all that is required for X to count as a believer/desirer. In his way, Dennett is a latter-day Rylean. What is required for mentality is the right pattern in the behaviour so we can see it as rationally appropriate. And just as S&J argued that Ryle has trouble in giving real explanatory clout to mental dispositions (conceived as he conceives them), so it might be argued that Dennett’s instrumentalism—if that is what it is—has difficulty in giving real explanatory clout to beliefs and desires (conceived as *he* conceives them).

Suppose we want to resist instrumentalism. Can we defend a realist line on beliefs and desires? A crucial dimension of assessment of a theory is its predictive power. The more correct predictions it gives, the better the chance we normally think it has of being true, with its postulated theoretical states existing. Astrology is an example of a bad or false theory (it yields repeatedly false or unreliable predictions of human behaviour on the basis of the positions of the planets and people’s time of birth). While, as we saw, Folk Psychology has great predictive power which we continually draw upon. So why not take it to be true?

Nonetheless, some philosophers maintain that the theory is, or is very likely to be, false. These are the **Eliminative Materialists** (‘eliminative’ because they then take it that talk of beliefs, desires should be eliminated in the search for true theories of persons, ‘materialist’ because it is assumed that the theory which should replace Folk Psychology should be a frankly materialist one).

According to Eliminative Materialists, then, there are no beliefs and desires. We have neurological states but in the true description of their operation, there will be nothing corresponding to beliefs or desires. There is nothing which satisfies the principles of Folk Psychology. The folk theory is simply false, not just slightly wrong in the occasional principle, but **radically false**. A new theory explaining our actions is needed. (Compare: there was once a folk theory that explained disturbed behaviour by reference to demonic possession. This folk theory wasn’t just slightly wrong: its ontology needs to be eliminated from an adequate theory of the world.)

Eliminative materialism certainly looks an unattractive position, given the apparent indispensability of Folk Psychology. Our reliance on Folk Psychology suggests that it is not something we should give up lightly, or junk along with astrology: so what arguments are there for eliminativism?

### 3. Arguments for Eliminative Materialism
A) Folk Psychology is plagued with failures. In particular it fails to provide any explanation of some central psychological features of humans; it leaves “an enormous backlog of anomalies and mysteries in its own explanatory domain.” E.g. why we need sleep; the nature of mental illness and the grounds of difference in intelligence. These show that it is “at best a highly superficial theory, a partial gloss on a deeper and more complex reality”.

But the fact that it cannot explain a certain range of phenomena need not undermine the truth of its account of others, in particular its everyday explanations of action. Why should we expect the theory to explain everything about us?

B) The theory is not the result of careful deliberation and theory formation in the way that our best scientific theories are. Rather it has been stagnant for several thousands of years. Good theories should develop their resources and extend the range of their successes.

It is not clear that the theory has been entirely stagnant for all that time. Take the explanation of action by reference to unconscious desire—any principles involving this would not have been currency before this century. Or take the rise of acknowledgement of the role of developmental situation (in addition to innate character) in explaining personality traits etc. Moreover we may just have been right all along: not all folk theories need fail.

C) Sometimes, when theory B succeeds theory A, we retain the ontology of theory A because we can find close enough reductive links between the ontology of B and the ontology of A (as when we still talk of ‘genes’ after the revolutionary discoveries of microbiology, because we can reduce talk of genes to talk of what’s encoded in our DNA). To retain mentalistic talk when we find the Great Neurobiological Theory To Come will require their to be a similar possibility of giving a reductive account of the mental in terms of the neurobiological. But, the Eliminative Materialist claims, the demands on such reduction (the demands on inter-theory relations when we can say that the old ontology is saved) are sufficiently strict and difficult to meet that it is implausible that our common-sense psychological theory would meet them.

The reduction of one theory, A, to another theory, B arguably requires there to be correspondence rules which explain how the entities referred to in A correspond to those in B. Additionally, the principles of A should follow from the theory B once their formulations are changed by replacing the terms of theory A with those which correspond with them. The reducing theory (B) in the present case would be the body of scientific theories including physics, organic chemistry, biology etc. The reduction of Folk Psychology would give correspondence rules linking beliefs, desires and other mental states to physical states (of the brain). And those brain states would stand in the required relations. But, the Eliminative Materialist argues, the prospect of this looks dim. ‘One is reminded of how the vitalist conception of life must have looked as organic chemistry marched forward.’(Churchland)

Is reduction in the strong sense really required? Can token-token identity theories hold a functionalist line whilst denying reduction? (Churchland appears to consider this response to be a cheat. It would be like saving alchemy and its alleged states such as ‘being ensouled by mercury’ by giving a functional characterisation of these. Is that a fair analogy?)

Eliminative materialists are not functionalists of the sort we have been discussing—since they deny we have beliefs a functionalist account of beliefs is of no use to them. Nonetheless, they can still choose whether or not they think the theory which will replace Folk Psychology (doing the job of predicting and explaining behaviour) should be a functionalist or not. The functionalist option (e.g. Stich) would advocate the functional individuation of mental states used in explaining behaviour etc., but these states would not be those referred to in Folk Psychology. The non-functionalist option (e.g. Churchland) will often expect Folk Psychology to be more radically replaced by neuroscience.

Reading
On instrumentalism: Are the beliefs and desires spoken of by folk psychology real states, or should we just treat belief/desire psychology as an instrument for generating reliable predictions?

Basics:

Further reading:

On Eliminative materialism: If we do take a realistic view of folk psychology, does that mean that the folk theory could be wrong, as much in need of elimination as—say—a folk theory of witchcraft?

Basics:
Terence Horgan and James Woodward: ‘Folk psychology is here to stay’, *Philosophical Review* (1985) 197–226; reprinted in M&C.

Further reading:
Jerry Fodor: ‘Banish DisContent’, in Lycan MC.

On the idea of ‘reduction’:
10. Sensations and Consciousness

1. A problem for functionalism?

Perceptual experience, it is natural to say, has a certain phenomenal character: for example, there is something that it is distinctively like to see a red expanse of colour or to taste Marmite or to hear a cello. Similarly with sensations: there is something it is distinctively like to feel toothache, or have an itch in the middle of your back.

Sometimes, the point is made using a bit of jargon: it is said that there are qualia\(^1\) distinctive of our various perceptual experiences or sensations, where qualia are features with a certain phenomenological character.

The functionalist, recall, identifies a mental state by its functional role. Can such an account be reconciled with the existence of qualia which are characterized by their phenomenological qualities (what they are like, rather than what they do)?

Arguably, not only does pain (for example) have a distinctive ‘feel’, i.e. there is something which it is distinctively like to be in pain, but this distinctive feel is what makes the experience a case of pain. In other words, arguably, what is characteristic of pain is a matter of qualia, not a matter of what pain causes and is caused by. And if this is right, then functionalism is in trouble as a general philosophy of mind. The opponent of functionalism who uses such an argument may allow that there are some mental states which are not associated with qualia (e.g. many beliefs may not have associated qualia: there is nothing it is distinctively like to believe that London is bigger than Sheffield), and such qualia-free states indeed may be functionally characterized. However, the argument runs, the functionalist story inevitably misses something important about perceptual experience and sensations, etc.

But is that right? Can the functionalist accommodate the intuition that there is something which it is distinctively like to be in pain, for example?

2. Functionalism and epiphenomenalism

Could someone have a brain with the normal human functional architecture busily doing its stuff in an entirely standard way, yet not be conscious and not be having experiences with a certain ‘feel’? If you answer ‘yes’, then you are treating consciousness and the experiencing of states with a distinctive phenomenology as—so to speak—an optional extra, a side-show perhaps accompanying typical human brain functions but perhaps not. On this view, therefore, a functionalist story about mental states being brain states interacting to produce behavioural output as a result of sensory stimulation etc. certainly cannot capture what is essential to consciousness because you can have the hierarchy of brain functions without the consciousness.

This idea that consciousness is an optional extra (an ‘epiphenomenon’ as the jargon has it) is, on reflection, pretty unattractive. After all, someone who shares our functional architecture will, by the contra-Cartesian hypothesis, behave exactly like you and me—because, of course, his behaviour is produced by a physical control system just like that which produces our own behaviour. Our functional equivalent will react to his environment, and initiate appropriate behaviour: he will be able to talk and answer questions, and even give sensible-sounding responses to our enquiries about what he is thinking about, or whether he is in pain, and so forth. In short, someone with the right functional architecture will certainly seem for all intents and purposes to be conscious; he will pass all the tests for consciousness that you and I pass. So what are we to make of the suggestion that, for all that, he might actually be a non-conscious zombie? And come to that, what is he going to make of the suggestion that he is really a zombie? Would you like to be the one to put to him that he could lack the je ne sais quoi of real consciousness?—being functionally structured just like you and me he is liable to object pretty vigorously!

There is no doubt some philosophical instruction to be had in trying to pinpoint exactly what is wrong with epiphenomenalism. But it is difficult to believe that there is any serious mileage in the idea that someone could be functionally equivalent to you and me (at all

---

\(^1\) Singular: quale [pronounced kwah-lay and certainly not “kwayle”!]
levels) and yet be entirely non-conscious. If you want to argue that pursuing the functionalist programme cannot illuminate the nature of conscious experience, the point will need to be made without backsliding into either Cartesianism or epiphenomenalism. How might the argument go?

3. What it is like

Let us consider the story of a man I will call Jack. Jack is born profoundly deaf; his condition is permanent and he never acquires the capacity to hear what is going on around him. Nor, thankfully, does he suffer from tinnitus; he simply has no conscious auditory sensations at all. Given his condition—we are strongly inclined to say—Jack does not know, and indeed cannot know, what it is like to hear. He does not know, and cannot know, what it is like to hear a sound with the timbre of a violin, or to hear a sound with the pitch of middle C as against top D. To know what these kinds of conscious experience are like, one must have had them (or at the very least, have had similar experiences); and by hypothesis, Jack completely lacks all experiences of the relevant auditory kinds.

Jack—to continue the story—eventually becomes interested in neuroscience: and he becomes especially interested in the neurophysiology of the human auditory system. Let’s suppose that he does first-rate work exploring the neural structures involved in human auditory perception, and he significantly advances our understanding of the hierarchical functional organization of the auditory cortex. Despite this expertise (we are still strongly inclined to say) there remains something which we all know but he cannot—namely what it is actually like, experientially speaking, to have auditory sensations. He may know all there is to know about the physiology of auditory perception, but if he doesn’t know what it is like to hear sounds, if he is unacquainted with the phenomenology of auditory sensation, then there is a sense in which he doesn’t know the most essential thing about the experience of hearing.

The moral of this story seems simple and compelling. It is one thing to know about the neurophysiological structures involved in some form of conscious sensation; it is something else entirely to know what it is like actually to have the sensation in question.

Thomas Nagel makes the same point by means of a slightly more exotic story in his justly famous paper ‘What is it like to be a bat?’ He asks us to consider those bats who find their way around the world by echo-location. As Nagel notes, ‘bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine’. In other words, this additional sense which the bats have and we lack presumably has its own distinctive character or phenomenology (as different from the phenomenology of vision and hearing as these are from each other); and this phenomenology is something with which we are not, and indeed cannot be, acquainted. Even the most expert investigator of the relevant neural structures in the echo-locating bats will remain as ignorant as the rest of us about what it is actually like to be a bat. Again the moral seems clear: to repeat, it is one thing to know about the objective neurophysiological structures involved in some form of sensation; it is something else entirely to know what it is subjectively like to have the sensation in question.

But this seems to indicate a fundamental shortcoming in any functionalist theory of the brain treated as an account of the mind. For we appear to have shown that a physical story about some kind of sensation—however detailed its account of our neurological functioning—must leave out an essential fact about the sensation, namely what it is like to have it. And hence any theory of the mind that tells us that sensations are just brain states with a certain functional role must fail dismally to give us a full account of the nature of sensation: work on what I called the Grand Unified Programme may increase our understanding of the way our brains function, but it cannot tell us about the subjective quality of conscious sensation. Note that the Nagelian is certainly not intending to resurrect the Cartesian idea that our behaviour is ultimately generated by a mysterious class of immaterial processes. Nor is he endorsing the epiphenomenalist idea that experiences may have the normal functional role yet lack a phenomenology. The idea is just that our experiences have a subjective phenomenal aspect as well as an objective one, and a functionalist theory only tells half the story.
4. Can the functionalist reply?

So we have a programme—try to understand consciousness by articulating the functional architecture of our cognitive processes. And we have an intuitively rather compelling claim—namely that no exploration of the functional structure of the mind/brain can deliver an account of what is essential to consciousness, namely the intrinsic phenomenological character of our experiences. Now, a very good motto in such circumstances is ignore the intuition and with any luck it will go away as you develop the programme. Do you intuit that simultaneity is absolute?—then immerse yourself in special relativity, and the intuition will begin to disappear. Do you find it intuitively compelling to think that every event must have a sufficient determining cause?—then a course in the delights of quantum mechanics should cure you. I’m inclined to add: learn to live with the functionalist programme and Nagelian intuitions will likewise go away. But I guess that that quick way with Nagel will be thought to be entirely disreputable and will win few converts; so I had better proceed more slowly.

Nagel relies on the premiss that there is a difference between knowing some functionalist theory about a given type of experience and knowing what that experience is like. I am happy to grant that this is true. But he slides from this premiss to a stronger claim: that there is some special kind of ‘phenomenological fact’ which we know when we know what the experience is like. It is this claim which, despite its intuitive appeal, I want to reject. For of course, once we accept this claim that there are phenomenological facts, the conclusion that there is a shortcoming in any functionalist theory about the experience would quickly follow—for the functionalist theory would fail to capture all the relevant facts. So what I need to argue is that the disputed claim about the existence of phenomenological facts doesn’t follow from the agreed premiss that there is more to knowing what an experience is like than knowing a functionalist theory about them.

But first, a clarificatory aside. Is the fundamental Nagelian claim that simply having an experience involves awareness of phenomenological facts, or is it that knowing what a certain experience is like requires acquaintance with the phenomenological facts? These are surely different claims, because having an experience is surely not the same thing as knowing reflectively what it is like. Consider Nagel’s bats again for a moment; they no doubt have experiences as they probe the world using their sonar sense; but I assume that real bats don’t go in for discursive reflective thought on their predicament—they are just not intellectually equipped for self-conscious reflection on their experiential state. I take it, then, that bats do not have any reflective knowledge about what their own sensations are like; although bats have experiences, they do not know what they are like in the way that we know what our experiences are like.

So is Nagel claiming that having a certain sort of experience is already a matter of being aware of some special ‘phenomenological fact’? Surely not. For this view—far from being intuitively appealing—is highly unattractive. Of course, as the bat flits around it is aware of perfectly ordinary facts (such as that there is an obstruction in front of it); but the suggestion now is that the sonar experience in virtue of which the bat is aware of the obstruction consists in awareness of another fact, a phenomenological one. This regressive multiplication of facts sounds depressingly like a updated version of the old philosophical mistake of thinking that awareness of outer objects consists of awareness of inner objects, mental images before the mind’s eye. I will assume without further ado that this is a dead end.

The Nagelian idea, I therefore take it, is not that experiences involve awareness of special facts, but rather that there are facts about our experiences which we know when we know what our experiences are like. And likewise there are phenomenological facts about the sonar experiences of bats, which we do not know, though the bats would if only they had the capacity for reflective self-consciousness. To counter this Nagelian idea, therefore, what I need to show is that knowing what our experiences are like need not be construed as a matter of being acquainted with some phenomenological fact—because, indeed, knowing what our experiences are like isn’t a case of factual knowledge at all.

To show this, let me spin again a tale which I used in The Philosophy of Mind, but which will (I hope) bear repetition. The tale concerns the history of Jill who—let’s imagine—grasps a functional theory about the processes which are involved when we experience a particular type of sensation which I will simply call ‘sensation S’ (this functional theory can be
articulated in as much detail as you like, with all the neurophysiological trimmings). At the outset of our tale, Jill can perhaps use her knowledge of the functional theory to attribute sensation $S$ to others, since the theory will tell her about any observable causes and effects of the processes which occur when one has sensation $S$. But she has yet to experience sensation $S$—or anything similar—for herself. So Jill is in the same relation to sensation $S$ as Jack was to auditory sensations: she only knows about $S$ from the outside, as it were. At this stage, then, it is natural to say (using Nagelian phraseology) that she doesn’t yet know what sensation $S$ is like, experientially speaking.

After a while, let us suppose, Jill does begin to have sensation $S$ from time to time. But, at least at first, she fails to recognize it as such: perhaps she fails even to notice that the same experience has recurred on a number of occasions. It continues to seem plausible to say that Jill hasn’t yet realized what sensation $S$ is like, experientially speaking.

Then one day the penny drops. Jill has sensation $S$ again. And this time, she says to herself - ‘Good heavens! It has never dawned on me before, but I seem to have that queer feeling in just the situations mentioned in the functionalist theory about sensation $S$. So that must be what sensation $S$ is like! I never paid it much attention before, but how distinctive it feels, now I come to concentrate on it!’ For a while, perhaps, Jill has to rely partly on clues from its typical causes and effects when she self-attributes sensation $S$. But after the occasional false start, Jill learns to identify straight off the occasions when she is having sensation $S$, without leaning on her functionalist theory. So at this final stage, Jill is in the same relation to her $S$ sensations as we are to our pains, for example. She recognizes them straight off; she knows from experience what they are like.

Now, at the beginning of this tale, Jill has a functionalist theory of sensation $S$ but doesn’t know what the sensation actually feels like; and at the end of the story, she does know what the sensation is like. What has happened in between? Well, two things. First, Jill started to have the sensation for herself; and then second, she learnt to identify sensation $S$ straight off, without relying on facts about its causes and effects. But having new sensations is not becoming aware of new information; and acquiring practical know-how is again not acquiring propositional knowledge. So the situation is this. It is quite true, as the Nagelian argument insists, that knowing a functionalist theory about sensation $S$ doesn’t give Jill what she needs if she is to count as knowing what sensation $S$ is like. But this is no shortcoming in the theory, since no theory could possibly give Jill what she needs. What she lacks at the beginning of her history isn’t more propositional information that could be encapsulated in an improved theory about sensation $S$—what she needs is the experience and know-how which no amount of information can provide. It is plainly no objection to a theory about sensation $S$ that knowing the facts stated in the theory won’t actually give you the sensation! And it is likewise no objection to a theory about sensation $S$ that understanding the theory won’t instantly enable you to say straight off when you are having the sensation. Learning to say when you have a particular sensation may take a bit of practice.

The crucial step in the Nagelian argument, remember, is the claim that what there is some special kind of fact which we know when we know what an experience is like. We can now see that this claim is resistible. Knowing what sensation $S$ is like is not grasping some special sort of phenomenological fact: knowing what sensation $S$ is like is being able to recognize it as such when you have it. Deprived of its crucial step, the Nagelian against functionalism fails.

5. Jackson’s Epiphenomenal Qualia and Inverted Spectra

A related argument is due to Frank Jackson. Imagine someone, Mary, who has always lived in a black and white room: she has never seen colours. She does not know what it is like to see red. What is it that she does not know? It is not physical facts, for we can imagine that she is an expert on, for example, what goes on in the brain when someone sees red, what things are red, what wavelengths of light are involved etc. Even if she knew all of the physical facts, she would still not know what it is like to see red. In short: having all of the physical information is not having all of the information.

So, the argument concludes, there are non-physical facts of which Mary ignorant: facts about the quality of the experience of seeing red. Supposedly, no materialist account can honour this conclusion: but the functionalist will essay a similar sort of reply as the reply to Nagel—knowing what it is like to see red is not knowing more information, but having experiences and being able to recognize them.
Still, is the functionalist off the hook yet? It seems that what it is like to see red does not effect one’s behaviour. Suppose Mary came out of her black and white room, it would make no difference to her subsequent behaviour whether green grass had for her what we would describe as a distinctively green look, or a distinctively red one. So, apparently, the exact character of the qualia she has will have no effect on behaviour, and more generally no effects on the physical world. They are epiphenomenal. They may be effects of what goes on in the brain, but they have no physical effects on the brain in turn. (On Jackson’s theory, mental states are, in general, causally efficacious; it is just certain properties of those states, namely qualia, which have no physical effects. Moreover, they may have other mental effects.)

Note, it is one thing to say—most implausibly—that consciousness generally is epiphenomenal (i.e. that being conscious is an optional extra, that may or may not be added to certain functional states); it is another thing to say that certain qualia are epiphenomenal. True, according to this account, qualia “do nothing, they explain nothing, they serve merely to soothe the intuitions of the dualists, and it is left a total mystery how they fit into the world view of science.” But Jackson replies that it is overly optimistic to assume that we can explain and understand how everything fits into the scheme of things. Epiphenomenal qualia may just be one unavoidable mystery.

To press the point that qualia may apparently escape the functionalist net: arguably it is a coherent possibility that when your sensation of seeing red is like my sensation of seeing green, and vice versa. This would mean that our qualia were inverted. Our behaviour would still match despite the difference in experienced subjective colours.

Can the functionalist allow this possibility? Your seeing red and my seeing red play the same functional role: they are both caused by red things and cause us to say ‘that is red’, so according to the functionalist they are the same mental state. This (allegedly) fails to account for the difference between our states on seeing red in the inverted spectrum case. Similarly, since your seeing green and my seeing red play different functional roles, the pure functionalist seems unable to account for the similarity which we are supposing there is. Other kinds of materialist can accommodate the idea of inverted spectra, however. For example, a species relative type-type identity theories may maintain that the mental type which is the sensation of seeing red is identical to that state which plays the relevant functional role in most members of the species. Perhaps you are typical as regards your sensations, whilst I am atypical: I have my wires crossed. But then, the sensation I get on seeing red things can count as a sensation of seeing green, because having green things in front of me causes me to go into the same brain state that most people go into when they see red things. So inverted qualia are not a problem for all materialists. (It might be tempting to press worries about whether the case of inverted spectra really make sense: are intersubjective comparisons of qualia possible? However an inverted spectrum within a person seems possible: you could wake up one morning and find roses look green and grass looked red. Supposing no-one else reported any change, the best explanation could be that your colour qualia had been inverted. So it seems to follow that you know that either before or after the switch, your qualia cannot be the same as those of other people.)

A related problem sometimes posed for the functionalist (and other materialists) arises from the claim that it is coherent to suppose that someone or something could be functionally equivalent to me (reacting in the same way to the same stimuli etc.) and yet, unlike me, not have any qualia. It could feel nothing at all. Take Ned Block’s example of the population of China acting out a person’s functional states. Surely the entity which is the population as a whole would have no qualia?

6. Functionalism again

A possible functionalist response to the qualia challenge would simply be to deny that qualia are relevant to the identity of mental states. As a consequence, they would not fall within the domain of psychology (or of cognitive science). This need not deny that some token mental states have qualitative properties, but this is no more essential to that mental state than the fact that the fact that the neurones whose states they are have, say a certain weight.

But does this do justice to the role of qualia in our mental life, and their centrality in conscious experience? Arguably it is that distinctive pain-quale which is essential to pain,
not its functional role. Qualia give us by far the most immediate and reliable knowledge of what sensations we are having - how can they be inessential to them?

But the property by which we can most effectively pick out something, need not be its defining property. We spot tigers by their stripes, but these are not essential to their tigerhood (there are albino tigers etc.). And is it plausible to suppose that pain feels just the same to other species?

On the other hand, some functionalists take the line that qualia are defined by their functional role. Hence anything with a state that fulfils that role automatically has qualia. (Then the Absent qualia story is not in fact possible.)

Note that qualia are properties of our states which enable us to discriminate certain sensations or experiences. Now, discrimination is behaviour which can figure in the functional characterization of, say, the sensation of red. So the properties, whatever they are, that enables the subject to so discriminate that sensation from others, can count as qualia (and these properties may be, for example, the presence or absence of certain chemicals in certain areas of the brain). These may be properties of our brain states for us: these may account for our introspective powers though we do not know (through introspecting) that it those properties which do so. For the robot, some other properties will play the comparable role.

The question of what qualia actually are, becomes a question for neuroscience not psychology.

7. The Relation between Qualia and Consciousness

Qualia are properties of mental states directly apprehended in consciousness. For there to be something that an experience is like, that experience must be a conscious one. There is nothing that it is like to be a stone because stones have no conscious states with qualitative feel. This has prompted some philosophers to claim that the having of qualia is what is distinctive of consciousness, and thus (for reasons like those already considered) the functionalist account of the mind misses out consciousness.

Although (ignoring problems with the notion of qualia in general) it may be plausible to say that it is only conscious mental states which have associated qualia, this does not guarantee that all conscious mental states have qualia. Conscious beliefs may not always have a distinctive qualitative feel, and other conscious intentional states may likewise lack qualia. The question of what makes a mental state conscious should not, therefore, coincide with questions about qualia.

On the other hand, an account of consciousness ought to explain the distinction between mental states which must be conscious to exist (e.g. you can’t have a non-conscious itch) and those which can be non-conscious. Does this coincide with the distinction between states with qualia (e.g. sensations) and states without?

8. Different senses of the notion of ‘consciousness’

We use the notion of ‘conscious’ in several different ways; let us clarify some of these.

We talk of someone being conscious in contrast with being asleep or anaesthetised. Consciousness, in this sense, may not be all-or-nothing nor clearly demarcated from the lack of consciousness. Some cases of sleep may count as more conscious than others. In particular, sleepwalking may be declared to have a conscious element (explaining the sleepwalker’s ability to avoid bumping into things).

Should we describe animals as being conscious in this sense? All animals? Just some? Where do you draw the line? This question might be of importance to issues such as vegetarianism. If we should not inflict unnecessary pain on conscious subjects, the question of whether they are conscious would be crucial.

We talk too of being conscious of sensations/having a conscious sensation. We are conscious, e.g. of our pains (query: does the notion if unconscious pain make sense?). The fact that pains are usually conscious might be explained by an evolutionary story: part of the function of pains is to intrude on our consciousness, thereby forcing us to draw our attention to the damaged area of the body. Compare having conscious beliefs (desires): most of our beliefs are not held consciously. No doubt you believe that zebras do not wear hats—but have you ever consciously entertained that thought before? And some of our beliefs and desires may be difficult (impossible?) to bring to consciousness.
The idea of perceptual consciousness (i.e. being conscious of the world around us) is closely linked to the ability to negotiate one’s surroundings. We talk of being conscious of the log in the road, for example. This is a different usage of ‘being conscious of’ from that the previous one—a sensation of belief which someone is conscious of is in some sense itself an element of that person’s consciousness in a way that the log clearly isn’t.

Can we also perceive non-consciously? Certainly we can walk around things without noticing we are doing it, that is without consciously perceiving that we are walking around anything at all. Is this perception without conscious sensations? (Compare also blindsight, and consider subliminal messages.)

What of being self-consciousness (not in the sense of being shy etc., but in the more general sense of being aware of oneself as a conscious being). Up to a certain age, babies might be said to be conscious of their surrounding, whilst not being conscious of their limited perspective—of their subjective self (e.g. they have not yet developed concepts of ‘my’ and ‘mine’). Once they pass this stage, they might be said to become self-conscious. The notion of self-consciousness is evidently tied up with knowledge of one’s own mental states, and hence with introspection (however that faculty is to be elucidated). But perhaps more is involved than atomistic awareness of this pain or that twinge: some sense of the bundle of mental events as hanging together in a

It has been suggested (see e.g. Wilkes) that the notion of consciousness is such a mess that we shouldn’t expect anything of an illuminating uniform account of all aspects of it. (Wilkes’ analogy: “no science explains what it is to be an ornament”). What kind of scientific discovery would be able to lay claim to having discovered the scientific basis of consciousness?

9. A Functionalist account of the consciousness of certain mental states?

Consider a Freudian unconscious desire—what is distinctive is that the subject does not believe (even denies) that she has that desire. Similarly for unconscious beliefs. Having the belief itself is a matter of behaving in certain ways (following the functionalist account of beliefs) but, when unconscious, there is no accompanying belief in its existence. Which suggests: conscious beliefs are those which are accompanied by beliefs in their own existence. That is to say, I consciously believe that p = 1 believe that p, and I believe that I believe that p. (The latter conjunct might be called a second-order belief: it is a belief about my beliefs.)

Similarly, if you adopt the account of perception as the acquisition of beliefs, then conscious perception will be the acquisition of conscious beliefs, i.e. beliefs accompanied by beliefs that one has acquired that belief. Unconscious perception involves acquiring information without realising it at the time (hence not having the second-order belief).

Likewise my conscious desire that p is my desire that p accompanied by a belief that I desire that p. And a sensation of mine is conscious if it is accompanied by a belief that I have that sensation. The fact that our sensations are nearly always conscious amounts to the fact that they are nearly always accompanied by a belief in their existence (a plausible suggestion). Indeed, there will be a causal connection between having a sensation and believing that one has that sensation (which may be part of the functional characterisation of that sensation). [This might account for the fact that the two states do not seem like separate states.]

The distinction between the first-order and second-order beliefs can be based on the behaviour associated with each (in accordance with the functionalist account of beliefs). In particular, the second-order belief is associated with the tendency to report that belief—hence conscious beliefs are the beliefs you are liable to report.

However …

Someone can have the second-order belief that they believe that p (i.e. behave appropriately) without consciously believing that p. E.g. Arnold might behave as if he believes that dualism is true (he is offended when anyone attacks it etc.) whilst not consciously believing that dualism is true (he denies that he believes it etc.). Conclusion: having the second-order belief is not sufficient for the consciousness of the first-order
belief. [Problems arise in fleshing out this objection with exactly what behaviour should be associated with the second-order belief in contrast with the first-order belief.]

2) The account mistakes the focus of attention of conscious believing. It implies that the conscious belief that pigs can fly is a belief about oneself (in particular about one’s beliefs), when in fact it is a belief about pigs in the world.

These two (non-conclusive) objections can be put forward by other functionalists (as they battle among themselves). An alternative account of consciousness, also acceptable to the functionalist, identifies conscious mental states by their being the actively, episodically thought about (as opposed to being the subject of dispositional belief states cf. Carruthers).

Objections which challenge this sort of higher-order belief/thought account more fundamentally will claim that it has entirely missed the essence of consciousness. ‘Consciousness is not a matter of having further beliefs or thoughts. Adding extra beliefs or thoughts will never capture the phenomenology of consciousness, or the fact that conscious states have characteristic feels.’ [Clearly this ties in very closely with the claim that functionalists cannot account for qualia.]

Reading

Background on ‘what it is like’:
S & J: Chapters 14 and 15.

Basic:

| O | D. Lewis: ‘What Experience Teaches’ in Lycan MC.

Further reading:
N. Block: ‘Troubles with Functionalism’, in Lycan MC. Also in Block RPP.
N. Block: ‘Are Absent Qualia Impossible?’, Philosophical Review 89 (1980) 257–74. (Reply to Shoemaker.)

Background on consciousness:
S & J: Chapter 15.
Churchland MC: Chapter 4.2

Basic:

Further reading:
W. Lycan: Consciousness, MIT 1987
11. Agency and Freedom

1. Actions and bodily happenings

Actions are things that you do, as opposed to things which happen to you: things which, to some extent, you have some choice about whether or not to do. They are up to you. Nervous tics or hiccoughs are not actions—you do not have control over them in the way that you (usually) have control over whether you do the washing up or clap your hands.

Of course, you may be coerced into acting in a certain way, as when a gun is pointed at your head: and you may say “I had no choice, I had to hand over the money...”. But this is an exaggeration: you did have the choice, even if an uninviting one, of not co-operating and getting shot. Handing over the money is an action which you chose to perform, even though perhaps no-one would think it sensible to make the contrary choice. Coercive cases don’t count against the thesis that actions involve choice, in some good sense.

Actions can be given different types of descriptions: some descriptions do little more than specify how I move my body, others specify the action in terms of various upshots and effects. The same happening could be described as raising my arm, or as voting, or as annoying Arnold, or causing an uprising... It may be all of these things. Not all of them may be things I intended to do. But, for it to be an action, the movement must be an intentional in some way or another (there must be a way of describing it as an intentional action, something I chose to do). Even if I didn’t intend to annoy Arnold, I intentionally did something which did annoy him.

The same pattern of overt body movements can constitute an action in one case but not in another. For example I may raise my arm, or it may be lifted by something else, and yet move in the same way. There is nothing distinctive about the mere movements which distinguishes actions from mere bodily happenings. So what does distinguish them?

Raising my arm differs from the case where my arm is lifted by other means because its cause is a mental state/event of mine. So can we define actions just as things you do that have a mental cause? No, this characterisation lets in non-actions. E.g. embarrassment (itself mental) can cause me to blush, without that being an action. An action has to result from the right kind of mental cause (something to do with choice and/or intention) in the right kind of way.

2. The Volitional Theory of Action

But what is it do something by choice or intentionally? Desires and beliefs will certainly play a role: e.g. my desire to vote for Zoë and my belief that now is the time to vote have something to do with me raising my hand. But are they enough? It might be said that a desire (and appropriate beliefs) is not enough—I must choose to act on the given desire. And this involves an act of will, via which my desire gets put into action.

The volitional theory of action states that the mental causes of actions are always acts of will, or volitions, and an action always has such a cause. Volitions can intervene and bridge the link between the beliefs and desires and our outward behaviour. But the crucial question is: what is a “volition” or “act of will”? In particular how are these “mental actions” to be distinguished from other mental happenings that we simply undergo (such as feeling embarrassment)? The answer seems to be that the acts of will are the mental events that happen through our choice. But to explain what happening through our choice consists in here, it seems that the volitionist’s initial reasoning implies that we would need to refer to a further act of will which causes the volition. A regress threatens (acts of will causing acts of will causing ...). Compare the sense-data theory of perception which explained perception in terms of seeing an internal, mental sense-datum but was threatened by a regress in explaining how the internal sense-datum is itself experienced. (For more discussion, see S&J, Ch. IX, §§1–5).

So: try a different tack. Acting by choice, intentionally, after all is just a matter of acting because of certain beliefs and desires which explain (“rationalize”) the action. So actions are those bodily happenings that spring from beliefs and desires in the right sort of way. See S&J, Chs IX §§6–8; XVII.

3. Causation and free action
But now, if we see actions as happenings brought about (causally) by beliefs and desires, where these beliefs and desires are themselves part of the causal nexus (according to the functionalist, they are token-identical to neural states, which are just part of the natural causal order), then in what sense can we be said to act freely? Is the thought that our actions are brought about causally by beliefs and desires (which are brought about causally by prior events) compatible with the idea that we sometimes act freely?

Why not? It might be said: the idea of acting freely is precisely the idea of our actions being brought about by our beliefs and desires (if our beliefs and desires didn’t make a causal difference, then we would be trapped inside our bodies, not freely moving them). And we want our beliefs and desires in turn to be causally constrained: we want our beliefs to be brought about e.g. by perceptual encounters or by rational chains of thought, we want our desires ultimately to be biologically appropriate. Causality is no threat to freedom. Even if the ultimate causal laws were sure-fire, non-chancy laws, still no threat. (Or so suggests S&J, Ch. XVIII)

But this happy compatibilism strikes many when they first encounter it as implausible. So in the rest of this part of the Outline,1 we will consider some considerations in favour of compatibilism (the aim is, so to speak, to get the position at least to the starting line for serious consideration).

4. Ayer’s argument

The headline idea: ‘incompatibilism is based on a confusion’—between causation and coercion/constraint

This is an indirect argument against incompatibilism (i.e. an argument which attempts to undermine the appeal of the position), which urges that the reason why people think that an action cannot be both freely performed and the result of antecedent causes is that they confuse causation with the more specialised notions of coercion and constraint:

... it is not, I think, causality that freedom is to be contrasted with, but constraint. And while it is true that being constrained to do an action entails being caused to do it, I shall try to show that the converse does not hold. I shall try to show that from the fact that my action is causally determined it does not necessarily follow that I am constrained to do it: and this is equivalent to saying that it does not necessarily follow that I am not free.

If I am constrained, I do not act freely. But in what circumstances can I legitimately be said to be constrained? An obvious instance is the case in which I am compelled by another person to do what he wants. In a case of this sort the compulsion need not be such as to deprive one of the power of choice. It is not required that the other person should have hypnotized me, or that he should make it physically impossible for me to go against his will. It is enough that he should induce me to do what he wants by making it clear to me that, if I do not, he will bring about some situation that I regard as even more undesirable than the consequences of the action that he wishes me to do. Thus, if the man points a pistol at my head I may still choose to disobey him; but this does not prevent its being true that if I do fall in with his wishes he can legitimately be said to have compelled me. And if the circumstances are such that no reasonable person would be expected to choose the other alternative, then the action that I am made to do is not one for which I am to be held morally responsible.


It is quite clear that many people do think that (1) the idea that we often act freely and (2) the idea that our actions are brought about in accordance with natural laws cannot both be true precisely because they refuse to accept that an action can be both freely performed and also entirely the result of antecedent causes. Furthermore, it may well be that their thinking this is the result of an erroneous generalization from the typical cases in which a causal factor has been brought to their attention in everyday life. For the causes of action to which they have

---

1 This will be very closely based on teaching material prepared by George Botterill.
been accustomed to pay heed may indeed involve causal interference with the agent’s normal choices and decisions, and as such may prevent the agent from acting freely or may make the action something for which the agent is not to be held (fully) responsible. But it does not follow from the fact that some causes frustrate freedom that all causes do so.

By way of analogy, imagine somebody who had only heard of bacteria in cases of bacterial infection. Such a person might well think that it is not possible for a healthy body to contain a large population of bacteria. Yet there are, of course, many species of bacteria that are in no way inimical to the health of the host, and indeed the presence of some types of bacteria is actually essential to the maintenance of life. Perhaps causes are to free action much as bacteria are to healthy life.

In general, there would seem to be three significantly different varieties of coercion or constraint, which we can label:

1. Total Agent-overriding Compulsion
2. External Outside-agent Compulsion
3. Internal Psychological Compulsion

We’ll discuss these in turn.

1] Total Agent-overriding Compulsion. Consider, e.g., the case of the Reluctant Dutch Boy. We have all heard the story of the heroic little Dutch boy who, seeing a breach in the dyke, realised that if nothing were done to dam the flow of water, there would be a disastrous flood. So he stuck his arm in the hole in the dyke. But what about the episode of the reluctant Dutch boy? This is a more sinister story. One day he was out walking with some companions when they spotted a breach in the dyke. There being nothing else to hand of a convenient size, his companions picked up the Dutch boy (who was certainly reluctant, because he put up quite a struggle) and jammed him into the hole in the dyke, holding him there while one of their number went for assistance in repairing the dyke. The reluctant Dutch boy stopped the incoming flow of sea-water just as effectively as the heroic Dutch boy. However, although his sufferings deserve sympathy, he is not to be praised for what he did— as the heroic Dutch boy was— since he did not do it of his own free will.

Or consider the case of the Shattered Vase. You are just showing me your priceless Ming vase, the pride of your collection. A deep-throated woofing is heard at the door and I turn to see your huge St. Bernard entering the room. It greets me affectionately, leaping up and placing a massive paw on each of my shoulders. Recoiling from the impact, I fall over backwards and crash down on top of the vase, smashing it to pieces. ‘Oh, my God’, you say, ‘just look what you’ve done. You have broken my vase.’ I have broken the vase. I can’t deny that, for some of the fragments are still stuck in my sweater. But I protest that I did not do it intentionally, and am not to be blamed for the accident.

What these cases seem to show is that we can be said to do things, even in cases in which our participation is that of patients, rather than agents. Some external agency may act upon us in such a way as to cause bodily movement, completely by-passing our psychological states. In such cases we do whatever our bodies do. But of course we do not act intentionally. As far as the events in question are concerned we might as well have been inanimate lumps of flesh. In such cases we are not to be held responsible for what happens, and we do not act freely because we do not act at all. I take it that this is unproblematic and agreed on all sides: compulsion that bypasses agency a fortiori excludes free action.

2] External Outside-agent Compulsion (the sort of case Ayer considers in the quoted passage). The case of coercion involving some threat deserves careful consideration. Ayer maintains that this is a case in which the agent is not acting freely and is not to be held morally responsible. Is he right about this?

At first, Ayer’s conclusions may sound very natural and tempting. But on reflection we can see that what he says about this sort of example is not entirely correct. Suppose it is a bank clerk who is forced at gun-point by an armed robber to hand over some money. The point that we can all (presumably) agree upon is that the clerk is not to be blamed for having done this. A manager who blamed the clerk for having done this would be more than stern and exacting. Such a manager would be downright unreasonable. But this is not because the clerk was not morally responsible for handing over the money. No, rather she is not to be
blamed because she did not do anything wrong. Faced with a choice of two evils (being killed or handing over the money), she quite rightly chose the lesser evil.

We can see that this is so by altering the case in such a way that the lesser evil becomes the greater evil. Imagine that what the armed raider demands is something more important than money—such as a quantity of nerve-gas sufficient to kill the whole population of the UK. Suppose further that the armed raider is a terrorist who is going to use that nerve-gas, and that since the person threatened is the only one present who knows the combination to the safe the terrorist will not be able to get it unless she hands it over. What do we say now? It may be understandable if you yield to such a threat even in those circumstances. (Which of us can be confident that she would do the right thing when threatened with death?) But, even so, such an action is one for which one can be held morally responsible, and what one ought to do in those circumstances is refuse to comply.

The clerk ‘had no choice’ and yet she ‘chose the lesser evil’. This sounds paradoxical, but isn’t. The explanation of the apparent paradox is that when we say that ‘we have no choice but to ...’, what we usually mean is that we do have a choice but that it is quite clear what choice we ought to make. ‘No choice’ is like ‘no contest’. What is ‘no contest’ is not something that is not a contest: it is rather a contest that is very one-sided.

Do we then agree with Ayer or not? The clerk acted intentionally, and rationally, in handing over the money. But she did not act of her own free will because she would prefer never to have been faced with that unpleasant choice. Moreover, she certainly did not hand over the money voluntarily or willingly. It is interesting to see how these classifications of actions peel apart, when applied to specific cases.

However, let’s not be distracted from the main point by too many details. Ayer’s main contention is that it is coercion and constraint in particular, rather than causation in general, that frustrates the exercise of free will. We have already seen that this contention appears to be borne out in the first category of cases (total agent-overriding coercion): there we do not have an exercise of free will because we do not have anything that can properly be called genuine action at all (even though these are things that you can be said to do, because your body does them). In the second category of case (external coercion by some other agent), although one can argue about the details of how these cases are to be described, Ayer’s main contention again seems to be defensible. These are unlike the first category of cases, because they are genuine actions. But they are not exercises of free will, because the agent is forced to do something which, in the absence of a threat, he would rather not have done.

Granting this, however, is not yet granting that Ayer has proved his case. For his main contention is a general one: that whenever an agent does something but does not act of his own free will, this is always because what he does is due to compulsion (coercion or constraint) rather than simply any form of causation.

3] Internal Psychological Compulsion. In the two previous kinds of cases the compulsion or coercion was external. It came from outside the agent and either caused the agent’s body to move in a way that totally by-passed the agent’s own psychological states, or else interfered with the agent’s own decision-making in an unwelcome way by making the agent do something that she would rather not have done, unless threatened with some dire consequence if she failed to comply. But can’t compulsion at least (if not coercion and duress) also be internal, deriving from some more or less chronic psychological condition within the agent?

This is what Ayer has to say about such cases:

A kleptomaniac is not a free agent, in respect of his stealing, because he does not go through any process of deciding whether or not to steal. Or rather, if he does go through such a process, it is irrelevant to his behaviour. Whatever he resolved to do, he would steal all the same. And it is this that distinguishes him from the ordinary thief.

But now it may be asked whether there is any essential difference between these cases and those in which the agent is commonly thought to be free. No doubt the ordinary thief does go through a process of deciding whether or not to steal, and no doubt it does affect his behaviour. If he resolved to refrain from stealing he could carry his resolution out. But if it be allowed that his making or not making this resolution is causally determined, then how can he be any more free than the kleptomaniac? It may be true that unlike the kleptomaniac he could refrain from stealing if he chose: but if
there is a cause, or set of causes, which necessitate his choosing as he does, how can he be said to have the power of choice? ...

The answer to this is that the cases I have mentioned as examples of constraint do differ from the others: and they differ in just the ways that I have tried to bring out. If I suffered from a compulsion neurosis, so that I got up and walked across the room, whether I wanted to or not, or if I did so because somebody else compelled me, then I should not be acting freely. But if I do it now, I shall be acting freely, just because these conditions do not obtain; and the fact that my action may nevertheless have a cause is, from this point of view, irrelevant. For it is not when my action has any cause at all, but only when it has a special sort of cause, that it is reckoned not to be free.


But has Ayer satisfactorily explained the difference between a kleptomaniac and an ordinary thief? He writes as if the kleptomaniac is trapped in a body which goes round stealing whatever he decides to do. But while that is a possible condition, it surely isn’t that of the real-life kleptomaniac (or the compulsive smoker, or whatever).

The compatibilist who claims that acting freely is to be contrasted with acting under coercion or compulsion ought to be able to explain the difference, because she supposes that an ordinary thief acts freely in stealing (and is therefore morally responsible for acts of theft committed), whereas a kleptomaniac (acting under compulsion) does not act freely and is therefore not morally responsible for such acts. But what is the force of ‘acting under compulsion’ here? Presumably there is some psychological state of the kleptomaniac that causes him or her to steal, when the opportunity arises. But won’t the same be true of an ‘ordinary’ thief? Indeed, won’t there have to be some internal state or other which causes the thief to steal? So what is the difference?

Admittedly, it is not just the compatibilist who would like to be able to explain the difference between ordinary behaviour and actions performed under psychological compulsion. Anybody who takes the view that there is such a thing as psychological compulsion is committed to there being some way of differentiating between individuals who are suffering from a compulsive condition and those who are not. But the difficulty is a particularly pressing one for the compatibilist who takes Ayer’s line that it is not causation in general, but coercion and compulsion in particular that freedom is to be contrasted with. Now, in cases of coercion and duress, when the agent is made to do something through being threatened by another party, it is quite clear that while these are no doubt cases in which causal processes are at work there are some quite special features that make such cases significantly different from ordinary causation. Above all, there is someone else who is interfering with your conduct by issuing a threat. It is by no means so clear what distinguishes behaviour under psychological compulsion from normal conduct. In both cases we have actions that result by some sort of causal process from previous psychological states of the agent. According to the compatibilist, it is not true that all causes compel. But what justifies calling one sort of case ‘compulsion’ and not the other?

Some possible suggestions for you to consider:

[1] One could take a hard and radical line by denying that there is any such thing as psychological compulsion, insisting that properly speaking compulsion is always coercion by others, and that talk of psychological compulsion is a mistaken attempt (excessively soft and liberal) to invent a new category of excuses. But I don’t think we could be very happy about that as a general stance. It is true that we should not assume that there must be a genuine condition corresponding to every label that a behavioural psychologist has ever devised (every ‘-pathy’, ‘-mania’, or ‘-phobia’). Repudiating all of these conditions is another matter. Nor is it true that the existence of some psychological compulsion is exclusively invoked in order to excuse conduct that would otherwise be morally blameworthy. The ‘compulsive behaviour’ associated with neuroses, for example, is often distressing to the agent, rather than harmful to anybody else (e.g., compulsive washing that the subject continues to indulge in even after her skin has become raw, blistered, and extremely painful). The wholesale rejection of the category of psychological compulsion also seems implausible in the light of what we know about post-hypnotic suggestion (which is, indeed, a sort of case that might be said to fall somewhere between external coercion and internal compulsion).
A more natural suggestion is that the difference between compulsive behaviour and normal behaviour is that compulsive behaviour fits into a general pattern that is recurrently and rigidly repeated. This seems quite a promising suggestion. The problem with it is that perhaps other behaviour also fits into a general pattern that we haven’t noticed yet. So maybe I am also subject to psychological compulsion in all my doings because I suffer from ‘normalitis’, a condition which afflicts all those who, having internalized the normal standards of conduct prevalent in the community in which they live, are then caused to act in accordance with those standards. Where’s the difference—except in the number of people exhibiting the different behavioural patterns, and the complexity of the patterns?

Perhaps the difference is that compulsive behaviour is irrational, whereas ordinary ‘normal’ behaviour must at least conform to some minimal standards of rationality. Until we have spelled out what those minimal standards of rationality might be, this proposal must remain rather too vague for any definite assessment. There is certainly something to be said for the proposal. For example, at least part of the difference between a kleptomaniac and an ordinary thief (to return to Ayer’s question) is that ordinary thieves do not steal if they feel that it is extremely likely they will be caught and punished, whereas kleptomaniacs seem to be undeterred (often? always?) by the likelihood of being apprehended.

5 Freedom is doing what you want to do

What appears to be a fairly simple consideration in favour of compatibilism may be presented as follows (to expand a little on a point made in §3). We may start by asking what anyone might possibly hope for in the way of freedom. What more could one expect in the way of freedom than the ability to do what one wants to do? Conversely, what is to be resented in any restriction upon one’s freedom is being unable to do what one wants to do. Could anyone ever aspire to a more perfect freedom in action than to do exactly what she wants to do, when she wants to do it? Such freedom, the compatibilist can now insist, is entirely consistent with our actions being part of the causal order. For why on earth should we think that it would be a consequence of the our actions being caused that nobody ever does what she wants to do? So long as actions are caused by our wants, what better?

Indeed, can’t we go further than this? In acting freely one does what one wants to do. Add to this the further assumption that when one does something because one wants to do it, the desire (one’s state of wanting to do that) is a cause of the bodily movement involved in the action (an assumption which seems natural and plausible). Will it not then follow, not only that acting freely is compatible with one’s action being caused, but also that in order to act freely one’s action must be caused —by some desire or other that one has?

A possible rejoinder to this compatibilist argument is that, so far from constituting a solution to the Free Will/Causation Problem, this point merely drives the problem inwards, from actions to their psychological antecedents. It might be argued:

It may seem as if it is sufficient for an action to be free that it should be caused by a desire. But if we bear in mind that desires too are caused, we will realize that desires (or, at any rate, events such as the onset of a desire, or a desire reaching a certain level of intensity) will be links in causal chains that stretch backwards in time, with earlier links that involve events external to the agent’s body and even prior to her existence. So, on this causal picture, an agent’s desires are causally dependent upon antecedent events and states over which the agent has no control. Therefore the agent would have no control over his desires, and hence no control over the actions that those desires produced and so be unfree.

Assessing the force of this rejoinder will require e.g. taking a hard look at the implicit assumption that if our basic desires are caused, then that is a Bad Thing. But is it? If my desires and beliefs are in some kind of harmony (I’m happy with my desires, unlike e.g. the kleptomaniac or compulsive craver after the drugs he knows are bad for him, i.e. I do not desire that certain of my desires cease, etc.) then why should I be worried by the thought that these desires are causally brought about in various ways? What I care about is being able to pursue these desires successfully, without hindrance, etc. But the matter is evidently a complex one. (See Dennett or Honderich for rich discussions.)
6. The argument against Contra-causal Freedom

If incompatibilism were true, then this would have to be because there was something in the concept of what it was to act freely that excluded the possibility of an action both being free and also caused to occur by antecedent events and states of affairs. In other words, freedom would have to be contra-causal. Many people feel that this is so, but fail to substantiate their conviction with any explicit account of what it is to act freely. Let’s consider what an explicit contra-causal definition of freedom would be like. Here is how one author has proposed to define what it is to be free with respect to an action:

If a person S is free with respect to a given action, then he is free to perform that action and free to refrain; no causal laws and antecedent conditions determine either that he will perform the action, or that he will not. It is within his power, at the time in question, to perform the action, and within his power to refrain. Consider the state U of the universe just before he takes or decides to take the action in question. If S is free with respect to the action in question, then it is causally or naturally possible both that U hold and S take (or decide to take) the action, and that U hold and S refrain from it.

(From Alvin Plantinga, *The Nature of Necessity*, pp. 163–4)

Of course, there is no question but that this sort of ‘contra-causal freedom’ is incompatible with our actions falling under strict causal laws: it is defined to be so. The question is whether we are ascribing this sort of freedom to an agent when we say that she does something freely, voluntarily, of her own free will. Posing this question introduces perhaps the strongest of arguments for compatibilism. For if we meant by ‘freedom contra-causal freedom, then in order to know that an agent had acted freely we would have to know that there were no antecedent causes of the action. In other words, we would have to know that it be possible, given whatever discovered or undiscovered laws of nature there are, that the whole history of the universe could have been the same and yet the agent could have acted differently. But we simply are not, never have been, and most probably never will be in possession of that information. So if freedom of action were contra-causal freedom, we could never be in a position to judge with any degree of confidence at all that an agent had, in performing some action, acted freely.

It seems, then, that an incompatibilist cannot know that his conception of freedom has any application whatsoever.

7. Responsibility

It is frequently suggested that accepting that the idea that our actions fall under causal laws is true would have very serious consequences for the way that we view and react to our fellow human beings. Indeed, it is often alleged that accepting that our actions are part of the natural causal nexus would (or should?) undermine our whole moral life; that we would no longer be able to regard people as being responsible for their own actions; and that we would therefore have to give up such practices as praising and blaming people for what they had done, and rewarding or punishing them according to their conduct.

The compatibilist may very well question whether these alarming results would really be the rational consequence of accepting the idea that our actions are part of the natural causal order. Perhaps if people were congenitally programmed to behave in a certain way throughout their lives, quite irrespective of how they were treated by others (like robots, and even rather crude robots without any degree of developmental plasticity), then it really would be pointless to praise and blame them, and of course, ex hypothesi, quite ineffective to reward or punish them. But the idea that our actions fall under natural laws is not the thesis that people are congenitally programmed and quite impervious to subsequent influence. It really would be alarming if that were true, but we have no reason at all to think that it is.

We can go further than that. We can surely assert with some confidence that, among the causal factors that influence (and perhaps collectively determine) how people behave, the way in which they are treated by other people plays a significant part. Approbation and reward encourage people to act in certain ways (they act as reinforcements of behaviour),
whereas blame and punishment dissuade and deter people from doing things or cause them to desist if they have already started.

Now, if this is granted, we may go on to ask whether accepting that our actions are part of the causal order really would be subversive of our moral thinking. Would it make morality pointless? There would still be actions which we would have every reason to encourage (because they enhanced utility) and others which we would have every reason to deter (because they caused suffering to others). In other words, so long as we recognised that the consequences of action could be desirable or undesirable by some moral standard, we would have a motive for promoting actions with good consequences and for discouraging actions with bad consequences. And we would still have reason to distinguish those cases where moral suasion can have some force (where ‘moral suasion’ is intended to cover such forms of inter-personal influence as moral argumentation—e.g., ‘you shouldn’t drink and drive, because you are thereby endangering other people’, ‘you shouldn’t smoke in public, because it puts other people, as well as yourself, at an increased risk of lung cancer’—positive inducement, threats of punishment, etc.). So we will still have reason to distinguish the sorts of cases where the agent can usefully be held responsible and moral suasion can get a grip.

It might still be said, however, that our received notion of responsibility somehow involves the idea of actions springing wholly from the originating person in a way that doesn’t fall under natural laws. Responsibility, it is sometimes said (cf. the quote from Plantinga), involves the ability to do otherwise, in a strong sense which implies that one could have done otherwise even had the causal setting stayed the same. It is very unclear that this is really part of our ordinary concept. But even if it is, then the suggestion will be that we should tidy up the received concept of responsibility into a new forward-looking, regulative concept of responsibility, according to which

X is to be held responsible for action A, if and only if A is an instance of a kind of action, K, such that X and other people can be influenced, by appropriate moral suasion, to act otherwise (i.e., to refrain from actions of kind K).

It is an interesting issue (though one that goes beyond our present concerns) how revisionary this definition is. Note though that it promises to be helpful in dealing with the distinction between kleptomaniacs and ordinary thieves that we previously considered in discussing Ayer’s argument for compatibilism. Kleptomaniacs are not responsible for their magpie inclination because their behaviour is not susceptible to modification in this respect either by moral reasoning or the threat of punishment. And note too that the revised concept of responsibility makes judgements of moral responsibility are partly evaluative in a way that seems correct. This is because whether people’s behaviour can be modified in various respects clearly depends upon the lengths to which you are prepared to go in order to modify it. Who can doubt that there would be far fewer kleptomaniacs, if the punishment for shoplifting were far more draconian - execution by beheading, or some process of torture, for example? Of course, we would not be prepared to go to those lengths. The deterrent would be inappropriately severe because the harm inflicted would be greater than the harm prevented. This is an evaluative judgement.

Let’s finish on responsibility with a (long) extract from J.J.C. Smart’s 1961 paper ‘Free Will, Praise and Blame’, *Mind* LXX, pp. 291-306:

Most of our ordinary senses of “could have” and “could not have” are not, in my view, incompatible with determinism. Though some of our ordinary talk about moral responsibility is frequently vitiated by a confused metaphysics of free will, much of it can be salvaged.

When in a moral context we say that a man could have or could not have done something we are concerned with the ascription of responsibility. What is it to ascribe responsibility? Suppose Tommy at school does not do his homework. If the schoolmaster thinks that this is because Tommy is really very stupid, then it is silly of him to abuse Tommy, to cane him or to threaten him. This would be sensible only if it were the case that this sort of treatment made stupid boys intelligent. With the possible exception of certain nineteenth century schoolmasters, no one has believed this. The schoolmaster says, then, that Tommy is not to blame, he just *could not* have done his
homework. Now suppose that the reason why Tommy did not do his homework is that he was lazy; perhaps he had just settled down to do it when some other boy tempted him to come out and climb a tree. In such a case the schoolmaster will hold Tommy responsible, and he will say that Tommy could have done his homework. By this he will not necessarily mean to deny that Tommy’s behaviour was the outcome of heredity and environment. The case is similar to that of the plate which could have broken.

The lazy boy is analogous to the china plate which could break and also could fall without breaking. The stupid boy is analogous to the aluminium plate: whatever the initial conditions the same thing happens. If Tommy is sufficiently stupid, then it does not matter whether he is exposed to temptation or not exposed to temptation, threatened or not threatened, cajoled or not cajoled. When his negligence is found out, he is not made less likely to repeat it by threats, promises, or punishments. On the other hand, the lazy boy can be influenced in such ways. Whether he does his homework or not is perhaps solely the outcome of environment, but one part of the environment is the threatening schoolmaster.

Threats and promises, punishments and rewards, the ascription of responsibility and the nonascription of responsibility, have therefore a clear pragmatic justification which is quite consistent with a wholehearted belief in metaphysical determinism. Indeed it implies a belief that our actions are very largely determined: if everything anyone did depended only on pure chance (i.e. if it depended on nothing) then threats and punishments would be quite ineffective. ... It begins to appear that the metaphysical question of determinism is quite irrelevant to the rationality of our ascription of responsibility.

What about praise and blame? These concepts are more difficult. We must at the outset distinguish two ways in which we commonly use the word “praise”. In one sense praise is the opposite of blame. We praise Tommy for his industry, blame him for his laziness. But when we praise a girl for her good looks this does not mean that we should have blamed her if her looks had been bad. When we praise one footballer for his brilliant run, we do not blame his unfortunate teammate who fumbled a pass. (Unless, of course, the fumble was due to carelessness.) When we praise Smith for his mathematical talent we do not imply that we blame Jones because, try as hard as he may, he cannot handle $x$’s and $y$’s. Of course we may well say that a girl is ugly, a footballer incompetent, or a man unmathematical, and this is the opposite of praise. But it is not blame. Praise and dispraise, in this sense, is simply grading a person as good or bad in some way. A young philosopher may feel pleasure at being praised by one of his eminent colleagues because he thereby knows that his work is assessed highly by one who is competent to judge, and he may be pained to hear himself dispraised because he thereby knows that his work is being assessed as of poor quality. Praise and dispraise of this sort has an obvious function just as has the grading of apples. A highly graded apple is bought and a highly graded philosopher is appointed to a lectureship, while a low graded apple is not bought and the low graded philosopher is not appointed.

In general to praise or dispraise a man, a woman’s nose, or a footballer’s style is to grade it, and if the grader is competent we feel sure that there are good reasons for the grading. In practice, of course, reasons are frequently given, and this giving of reasons in itself can constitute what is called praise or dispraise. For example, if a philosopher writes about some candidate for a lectureship that he has some illuminating new ideas about the logic of certain psychological concepts, this is the sort of thing that is meant by “praise”, and if he says that the candidate is muddle-headed and incapable of writing clear prose, this is the sort of thing that is meant by “dispraise”. It is not the sort of thing we mean when we contrast praise with blame. To say that a man cannot write clear prose is not necessarily to blame him. He may have been brought up among muddle-headed people and always given muddle-headed books to read. The fact that we do not feel like blaming him, however, does not alter the fact that we warn prospective employers about him.

Just as we may praise or dispraise a woman for her figure, a footballer for his fleetness or slowness of foot, a lecturer in philosophy for his intelligence or lack of intelligence, and a writer for his clarity or obscurity, so naturally enough, we may
praise or dispraise a man for his honesty or dishonesty, truthfulness or untruthfulness, kindness or unkindness and so on. In this sense of “praise” we may praise moral qualities and moral actions in exactly the same way as we praise beauty, intelligence, agility, or strength. Either we may do so quite generally, using a grading word like “good”, “excellent”, or “first-class”, or we may simply give a description. ... Praise has a primary function and a secondary function. In its primary function it is just to tell people what people are like. To say that one candidate for a lectureship writes clear prose whereas another cannot put a decent sentence together is to help the committee to decide who should be given the lectureship. Naturally enough, therefore, we like to be praised, hate to be dispraised. And even if no actual advantage is to come from praise, we like to be praised by a competent judge for work we have done because we take this as evidence that we have been on the right track and done something valuable. Because we come to like being praised and to hate being dispraised, praise and dispraise come to have an important secondary function. To praise a class of actions is to encourage people to do actions of that class. And utility of an action normally, but not always, corresponds to utility of praise of it.

So far I have talked of praise and dispraise, not of praise and blame. This is because I wanted a contrary for “praise” in the sense in which we can praise not only a moral action but a woman’s nose. What about the contrast of praise with blame? Here I suggest that a clear-headed man will use the word “praise” just as before, and the word “blame” just like the previous “dispraise”, with one proviso. This is that to praise (in this sense) or to blame a person for an action is not only to grade it (morally) but to imply that it is something for which the person is responsible, in the perfectly ordinary and nonmetaphysical sense of “responsible” which we have analysed earlier in this article. So we blame Tommy for his bad homework if this is due to laziness, not if it is due to stupidity. Blame in this sense can be just as dispassionate as dispraise of a woman’s nose: it is just a grading plus an ascription of responsibility. It is perfectly compatible with a recognition that the lazy Tommy is what he is simply as a result of heredity plus environment (and perhaps pure chance).

Now most men do not, in my opinion, praise and blame people in this dispassionate and clear-headed way. ... most men do not feel that blame, in the way they use the word “blame”, would be appropriate if a man’s action was the result of heredity plus environment. The appropriateness of praise and blame is bound up, in the eyes of the ordinary man, with a notion of free will that is quite metaphysical. Admittedly this metaphysics is incoherent and unformulated (as indeed it has to be, for when formulated it becomes self-contradictory). Nevertheless we can see that a rather pharisaical attitude to sinners and an almost equally unhealthy attitude to saints is bound up with this metaphysics in the thinking of the ordinary man if we look at the way in which very often his whole outlook and tendency to judge (not just to grade) other men changes when he is introduced to, and becomes convinced by, a philosophical analysis of free will like the one in the present paper. ...

The upshot of the discussion is that we should be quite as ready to grade a person for his moral qualities as for his nonmoral qualities, but we should stop judging him. (Unless “judge” just means “grade”, as in “judging apples”.) Moreover, if blame in general is irrational, so must be self-blame or self-reproach, unless this comes simply to resolving to do better next time.

8. Concluding note

Philosophical beginners often find it puzzling to be told that probably most contemporary serious writers on free action think that the idea that we often act freely is compatible with the idea that our actions are part of the natural causal order (even if the natural laws are sure-fire, deterministic, ones). The aim of the recent sections has not been to establish compatibilism as true, but to establish it as a contender—a position meritings serious consideration. So we have not tried for a comprehensive overview of the area, just to counterbalance the beginner’s typical prejudice against compatibilism.
Essential background:
S & J: Chapters IX, XVII and XVIII.

On actions and their causation:

On free will
12. Thinking and the Language of Thought

1. Thinking: Preamble

Thinking, in the sense that now concerns us, is an occurrent mental process. The occurrence of an individual thought is a particular event: thinking involves engaging in a series of such occurrences which are connected to one another in particular ways that can vary between rational reasoning and idle day-dreaming. (Note: sometimes, talk about what a person thinks is just talk about her beliefs, which can be non-occurrent. You doubtless think that baboons don’t wear pyjamas, although until you read this sentence, the thought had probably not occurred to you.)

Some occurrent thinking is conscious and involves processes over which you have some control (unlike, for example, the lack of control over what you are perceiving). We can call this deliberative, discursive thought. This will include thinking through what to say to someone, or working out how many lectures you have been to, or carrying out other processes of reasoning. Some such thinking may be taken to be a form of ‘inner verbalisation’: you have those words ‘going through your head’ that you would use to express your thoughts in conversation. Other conscious thinking may involve visualising images: I might think through how to arrange a room by imagining how it would look with various furniture in different places. Similarly, a composer may do some of her thinking by running through melodies in her head.

More controversially, non-conscious thought processes also seem possible. I might come up with the answer to a puzzle after a period of time in which I have not been concentrating my attention—arguably this was through an unconscious thought-process. Similarly, there may be complex actions or activities which require thought, but which can be done without conscious attention.

Occurrent thoughts (just like beliefs, desires, intentions etc.) have content—you think that p. Moreover, thoughts can have the same contents as beliefs and desires. So the same questions arise about the intentionality of thoughts as about the intentionality of other mental states. E.g. how do they have content at all, and why does a given thought-episode have the content it does?

For more introductory distinctions and elementary discussion, see S&J Ch. XVI. This isn’t the most wonderfully constructed chapter in that book. But you should take away from that discussion arguments for the following two related headline points:

1) The Ryle/Wittgenstein insight that doing something thoughtfully (whether playing tennis, calculating, or soliloquizing) isn’t doing two things at once—as it might be, playing tennis and thinking—but doing one thing in an intelligent, attentive, purposive way. In particular, what makes ‘thinking aloud’ a case of thinking isn’t that it is accompanied by some silent commentary which is the real thinking. In other words, what makes discursive overt thinking and covert thinking in one’s head both cases of thought is not that the first somehow always involves the second, but that the language in either case is being used overtly in an intelligent, attentive, purposive way. (“Expressing a thought” needn’t be putting into words an independently existing item which is the real thought.)

2) We should aim to give an account of covert thinking by (a) giving a story about thinking aloud, and (b) giving a general story about ‘doing things in one’s head’: factoring together these stories will deliver the desired story about thinking to oneself.

The second point, in particular, stands opposed to a rather different line on the order of explanation, which is the topic of the rest of these notes on thinking.

2. The Language of Thought Hypothesis

According to the Language of Thought (LOT) hypothesis, all thought occurs in a language. But rather than this being a natural language (English or Russian etc.), it is a Language of
Thought. (It is clear that not all thought occurs in natural language—the chess player’s thinking may consist in manipulating images of a board, the musician’s thinking in manipulating auditory representations: and perhaps some people think though they have no public language at all.) This Language of Thought is given the name Mentalese.

Just as English has symbols, namely words, which can be combined in various ways to form sentences, so Mentalese is taken to have its own symbols which combine to form sentences in the language of thought. And just as words have meanings and the meaning of sentences somehow derives from them, the same goes with the symbols and sentences of Mentalese. **When you have a thought that p, you have in your head a token of a sentence in Mentalese which means that p.**

Tokening a sentence of Mentalese is sometimes said to be rather like having an inscription of a sentence on a blackboard in your head, except the tokens of the symbols of Mentalese are usually taken to be neural objects (i.e. materialistically acceptable candidates). But such analogies shouldn’t be taken too seriously: blackboards need intelligent, thoughtful readers, and we don’t want a regress of thinkers.

In the classic version propounded by Jerry Fodor, it isn’t just current thoughts that involve Mentalese sentences. Beliefs and other propositional attitudes can get their content through the meanings of symbols in Mentalese. For example: **To believe that p is to bear a certain relation to a token of a symbol which means that p.** Desiring that p would be bearing a different relation to a token of the same symbol.

Beliefs, desires and current thoughts are structured states. The thoughts that “ducks can fly” and “Peter is troubled by ducks” have in common a symbol which means “ducks”. Thus, not only are the contents of intentional states complex (something most people agree about), but the states themselves are too. Just as sentences are made up of words which can appear in different sentences, so the claim is that thoughts too have constituents (symbols in the language of thought), and the same symbol can be a constituent in various different thoughts. Similarly, just as some sentences (e.g. “I’m going for a run and I’m going to take a shower”) have whole sentences as constituents (the sentence “I’m going for a run” and the sentence “I’m going to take a shower”), so thoughts can have thoughts as constituents (e.g. the thought that I’m going for a run and I’m going to take a shower has as a constituent the thought that I’m going for a run).

People who deny the LOT hypothesis may object to this. They may agree that token thoughts are states of the brain (if they are materialists), but deny that they are sentence-like with shareable constituents.

Since tokens of mentalese sentences are physical, we can happily take them to have causes and effects. Thus, for example, suppose I intend to stamp my foot. This state is realised by my relation to a token sentence in my head which means (in Mentalese) “I stamp my foot”. This state of my brain ‘churns and gurgles and computes and causes’, with the effect that I stamp my foot. Similarly, during a thought process, one symbol tokening will cause the tokening of a different symbol, for example producing the thought “this banana is yellow” from a thought “every banana is yellow and this is a banana”.

### 3. Computation and the Mind

Suppose we accept the LOT hypothesis (more on reasons for this below). A comparison with computers suggests itself once again. Thought processes, manipulating Mentalese representations, can be compared to a computer’s manipulation of symbols.

Take your reasoning from the thought that every banana is yellow and the thought that this is a banana, to your thought “this is yellow”. You have the first two thoughts by having symbols in your head with the corresponding meanings. There is a mechanism (specified by an algorithm: instructions by which your brain works) which surveys those symbols, processes them in a specified way and comes up with a string of symbols meaning that this banana is yellow.

You can programme a computer to produce a sentence of the form “this is Y” from sentences of the form “every B is Y” and “this is a B” whatever words are put in for Y and B. The computer need not understand the words, nor why the rule is a good one’ it can follow the rules mechanically and ignorantly. All that it needs is to be able to recognise physical
features of the symbols which appear in the sentences, and to be able to reproduce symbols for its output according to the rule.

Similarly, (the LOT hypothesis can say) thought occurs through comparable mechanisms which involve surveying the Mentalese symbols and this causing the tokening of another Mentalese sentence according to rules which can be followed mechanically.

Such mechanical rules can be immensely powerful. A small number of rules can generate a vast range of reasoning (compare instructions for doing long multiplication which can be used to calculate an infinite number of sums). Note that the rules governing our thought processes will be complex, but the same principles apply.

To reiterate: a computer can compute without knowing the meaning of symbols. Instructions don’t need to be implemented by intelligent agents, they can be realized as a set of automatic computations. But, (with a complex enough instructions), you can get out of the computer apparently intelligent trains of thought, if, for example, the mechanism always produces a thought which logically follows from previous ones. Similarly, the mechanisms by which our thought processes work need not themselves be intelligent for us to manifest intelligence.

It might seem that a sensible thought process, involving the manipulation of sentences of the language of thought would need an intelligent inner agent choosing one sentence after another. Thinking would need thoughtful choice of successive states; the thought needed for that choice would require further thought ... etc. A regress would ensue. The computational view of the mind stops the regress by denying that the choice of the next state needs thought. It can be picked by processes that involve no intelligence, in particular through rules which mechanically pick the next state according to the physical properties of the previous one. The Language of Thought hypothesis allows thoughts to be symbols with properties conducive to this type of account.

How do symbols in Mentalese get their meaning? This is a question about intentionality that we looked at in Lecture Topic 2. There is a symbol which is tokened when I have thoughts about ducks: in virtue of what does this symbol represent ducks? What is the relation between symbols and the things they stand for? A theory of content is needed. But note that since other propositional attitudes are taken to be relations to symbols in Mentalese, we only need do the job once and account for the content of symbols of Mentalese.

The meaning of natural language expressions is conventional: it is an arbitrary social matter that “cat” refers to cats and not dogs. Mentalese expressions must have their references through a natural connection between the Mentalese word for “cat” and real cats in the world: it cannot be through human intervention. What kind of account can be given for that connection? Some possible answers were considered in an earlier lecture on Intentionality (in particular resemblance, reliable indicator and teleological accounts).

Note that the above account of thinking puts the question of content to one side. A computational account of thinking explains thought processes in terms of symbol manipulation without drawing on what the symbols mean.

We can consider the rules which generate our thought processes by operations on symbols independently of both the meanings of those symbols and independently of describing the physical system which implements the rule. Such study of the mind falls within the domain of Cognitive Science.

4. Why believe in a Language of Thought?

We have two hypotheses which have, so far, been rather run together:

1. occurrent thought takes place in a LOT, Mentalese
2. propositional attitudes also involve processing tokens of Mentalese.

The story is that, when thinking, we have systems of physical states which represent and have content in much the way that written natural language sentences do. I.e. some physical states will serve as elements of the vocabulary and refer to things and properties in the world (these will be concepts: compare words in a natural language). And they will be combined according to (grammatical) rules to form complex representations or “sentences” with meanings determined by the references of those parts. Those representations will account for the content of our thoughts and other intentional states. Those contents also stand in logical
relations to one another. Thought processes can occur by manipulation of the (physically realised) symbols according to rules which operate on the basis of their physical features rather than their content. But a co-ordination between these rules and logical relations between the thought contents assures that the resulting thought is intelligent.

Note that we can in principle unpick the ingredients of this package. We could affirm [1] while denying [2]. The picture then would be that we have beliefs and desires, for which a Mentalese-independent account of content can be given, and then we have a special module for occurrent thought which encodes belief contents into Mentalese sentences, shuffles these about to infer other sentences, which then are decoded back into belief contents. It is unclear, however, why anyone should want to endorse this picture, however. For isn’t it at least as plausible, and more economical, to say that we have a special module for occurrent thought which normally encodes belief contents into natural language sentences, shuffles these about to infer other sentences, which then are decoded back into belief contents?—after all, most of our internal discourse with ourselves, our discursive thought, seems to be in English, or whatever. In other words, if [2] is rejected, it is natural to say that occurrent thought typically occurs directly in our natural language (though some thought involves e.g. the manipulation of images—the chess player running through future positions). What makes these uses of language count as cases of thought is their intelligent purposiveness (Ryle again): and the uses of language get their content in virtue of their conventional use to express beliefs whose content is not linguistic.

Rather more plausibly, we might affirm [2] and deny [1]. That is to say, we might agree that in order to account for beliefs we need to think of them as relations to sentences of a LOT. But we might argue, as just suggested, that occurrent discursive thought involves processing these sentences into natural language. Though then it is puzzling why we don’t think in LOT if we have those representations available.

The distinctive claim of the LOT hypothesis is that thoughts, beliefs, desires etc. all have constituent structure, as sentences do. Why accept that? Why think that having a thought is having a token of a sentence-like entity in your head?

Very sketchy remarks in favour of the Language of Thought will draw on comparisons with computers: just as they employ “machine languages”, so it is reasonable to posit human “machine languages” in which we think. Slightly less sketchy arguments are as follows:

(1) **Productivity.** The LOT hypothesis helps answer the question of why there are so many (indefinitely many) propositional attitudes or thoughts that we can have—a huge stock encompassing thoughts of almost indefinite complexity. This feature is known as productivity. This is comparable to the productivity of language—there are indefinitely many sentences you could utter and sentences can be indefinitely complex. This feature of language is explicable by the way sentences are made up of smaller units (words). Rules by which words can be combined to form sentences allow the construction of indefinitely many and indefinitely complex sentences. If propositional attitudes are relations to sentences in a language of thought, then the explanation can be of the same form. (So the argument from productivity is an “argument from the best explanation”.)

(2) **Systematicity.** You don’t get speakers of English who can understand “John loves Mary”, but not “Mary loves John”. More generally: the ability understand some sentences is intrinsically connected to the ability to understand many of the others. This is because learning English involves learning the way the meaning of a sentence is determined by the meanings of its words, so the way sentences are built up from smaller units (words) is crucial. In the above example, all the same words appear in the two sentences, so you’ll understand both or neither. But this systematicity goes for thought too: whoever is able to think (or believe) that John loves Mary can think that Mary loves John. The best explanation of this fact is that thought occurs in a language which similarly has constituent structure.

(3) **The constituency of behaviour.** The fact that a certain cause c₁ has effect e₁ and cause c₂ has effect e₂ and cause c₃ has both the effects e₁ and e₂, is reason to think that c₃ has c₁ and c₂ as constituents. Now, suppose c₁ is the intention to raise my left hand which causes the effect, e₁, the raising of my left hand. Whilst c₂ is the intention to hop on my right foot which causes e₂, hopping on my right foot. Consider the cause c₃: the intention to raise my left hand and
hop on my right foot, this has both effects e₁ and e₂. LOT hypothesis claims that the best explanation is that this intention, c₃, has c₁ and c₂ as constituents. Thus this argument is of the following form: behaviour has constituent structure, hence its mental causes must have too.

(4) Thought processes. A fourth (and central) argument emphasises the success of the non-regressive explanation of thought processes in terms of symbol manipulation (discussed above): an explanation which the LOT hypothesis makes available.

Queries: (1) how do these arguments relate to components [1] and [2]. What responses might be made to these considerations? Is it fair to complain that these arguments wrongly read features of the way we ascribe beliefs (i.e. features of the content-sentences we use, which have constituent structure etc.) into the states which they are used to ascribe?

(2) The LOT hypothesis seeks to explain contents of beliefs and other propositional attitudes in terms of the meanings of expressions in a language (he language of thought). But can we make any sense of the meaning of expressions in a language, without first assigning contents to the beliefs and intentions of its users? LOT arguably puts the cart before the horse. Surely the content of beliefs give content to the symbols in the occurrent thoughts (overt or covert) that express them, rather than vice versa. Or is this to misunderstand the LOT hypothesis?

(3) Do we all think in the same language of thought? Is all thinking in Mentalese? Certainly we seem to think in English sometimes—when we have a train of natural language sentences going through our heads (inner verbalisation). But someone who thought we always think in Mentalese could maintain that the genuine thought is going on in Mentalese, it just gets translated into natural language in our heads. (And when we can’t find the right words to say what we are thinking, the problem is at the translation stage.) But is this right? Cf. the Ryle/Wittgenstein critique of the idea that thought in natural language requires backing with some other kind of thought?

Reading

S & J: Chapter 16.
C. McGinn The Character of Mind: Chapter 4.
Churchland MC: Chapter 6 (On A.I.)


D. Dennett: ‘Styles of mental representation’ in Dennett’s The Intentional Stance (M.I.T. 1987) 213–36
13. Representational Content

Beliefs and desires have content. In believing that the coffee is ready, I take the world as being a certain way. Similarly, in desiring that I have a cup of coffee, I want the world to be a certain way. So beliefs and desires have truth/satisfaction conditions – i.e. there are worldly conditions which would need to be fulfilled for the belief to be true, or the desire to be satisfied, there are (actual or possible) states of affairs which the beliefs and desires are about.

This feature of ‘aboutness’ is standardly called intentionality and beliefs, desires and the other propositional attitudes are called intentional states (the label is inherited from medieval logic). Note, ‘intentional’ in this special sense – possessing aboutness—is not to be confused with ‘intentional’ meaning ‘purposive’. Intentions (such as my intention that I lose weight) have aboutness, and so are intentional states in the technical sense; but of course not all intentional states (i.e. states with propositional content) are intentions.

The child’s belief that Father Christmas delivers the presents is about Father Christmas. Sir Galahad wants to discover the Fountain of Youth – his desire is about that Fountain. Which illustrates that the things that my beliefs and desires are about need not exist. Another example: if I have beliefs about God, it doesn’t follow that he exists. Contrast this with ordinary relations (e.g. kissing) which cannot obtain without the existence of both of the things related. Little Johnny cannot kiss Father Christmas if Father Christmas doesn’t exist. But he can have a belief about Father Christmas without him existing. (This contrast might seem to indicate that there will be a problem giving an account of intentionality in terms of physical relations.)

Brentano claimed that intentionality was the mark of the mental. But pains seem to be mental without displaying intentionality—so it can at most be a sufficient condition of the mental.

Non-mental Representations

Beliefs, desires and other propositional attitudes are often described as representational states. They represent possible states of affairs. But don’t mistake this alternative label for a theory. We might be say e.g. that ‘Beliefs are like mental maps of the world.’ But instances of representation other than mental states (like maps, or representations in art or language) seem to inherit their power to represent from us. A series of lines on a piece of paper represents the layout of the London Underground, or a stormy sea, because we interpret the lines in that way. Similarly a string of noises issuing from my lips represents the fact that it is raining in virtue of the interpretation of these words by me and other users of my language. These other cases of representation will be of no help in understanding how our own mental states represent, if they themselves depend on mental representation.

Two Questions Distinguished

(MQ) What is the metaphysics of intentional states? (Are intentional states causal: if so, how do they relate to the causal relevant brain-states? Can a belief-state be identified e.g. with a brain state in the way that lightning [folk notion] is identified with a discharge of electricity [scientific notion]? Should intentional states appear in our ultimate catalogue of what there is, or are they merely a helpful fiction? – cf. centres of gravity). §§5, 8, 9 of these notes are relevant.

(SQ) How can one state (‘in the head’, maybe a state to be identified with the state of some clump of neurones) represent something else (almost always extra-cranial)? What it is for e.g. a belief to represent a particular state of affairs? In virtue of what do we count as having a belief with one particular content rather than another? This might be dubbed the semantic question.

Answering (MQ) needn’t answer (SQ). E.g. a type-identity theorist has a line on the metaphysics of belief which doesn’t in itself tell us why beliefs have the contents they do.

The Naturalistic Constraint
We want an account of intentionality which enables us to see how intentionality fits into the natural physical world. It surely isn’t just a brute, incomprehensible, datum that some things are ‘about’ other things. That is to say, aboutness will surely not appear in the complete catalogue of the ultimate and irreducible properties of things (aboutness isn’t like mass or electrical charge). A tendentious but striking way of putting the point: ‘If aboutness is real, it must be really something else.’ (Fodor) So we want a theory which explains how aboutness arises in the natural world.

(MQ) Relations to Propositions?

To see one way in which the Naturalistic Constraint kicks in, consider the suggestion that propositional attitudes are genuinely relational: when Peter believes that Zoë is clever, that is in virtue of the belief-relation holding between Peter and the proposition that Zoë is clever.

Propositions are standardly taken to be abstract entities (as are e.g. numbers)—they are not located in space and time, and do not have causal interactions with things that are. (On one view, propositions can be taken to be sets of possible worlds: on another view, propositions are structured set-theoretic entities—but they are non-physical either way.) How can a relation involving an abstract object have causal powers? If beliefs are relations to abstract propositions, how can they be causally efficacious? That looks problematic given the Naturalistic Constraint that we integrate our understanding of the role of contentful states into our best going story of the physical world.

Response (1): the apparent relational aspect of ‘Peter believes that Zoë is clever’ is an illusion. Beliefs are monadic states of people: e.g. they are states which play a certain functional role in interacting with other intentional states and producing behaviour. A comparable case?—physical properties such as heights can be described as relations to numbers. E.g. someone can be said to stand in the relation ‘height-in-feet’ to the number 6. But such physical properties are not genuine relations—they are monadic physical properties. And just as being-six-foot is a monadic property, so is believing-that-Zoë-is-clever.

Response (2): We can after all take propositional attitudes to be relational without taking them to be relations to propositions by taking them to be relations to mental symbols or ‘sentences in a language of thought’. So, believing that grass is green is a matter of being related in a certain way to a mental item meaning ‘grass being green’. More on this in Special Topic D. But NB this doesn’t begin to answer (SQ)—for an account of mental representation within the ‘language of thought’ is still required—why does a certain symbol represent (or mean) one thing rather than another?

(SQ) Representation and Similarity

Does mental representation work by similarity between the representation and the thing (or state of affairs) represented. It might be said that a picture represents a cat because it is similar to a cat in visual respects, and likewise my idea of a cat is similarly similar to the real thing.

1) There are general difficulties with determining what counts as similarity. If you count shared properties this will be no guide because any two things share many, many properties (if often very artificial ones). And is my two-dimensional paper picture of a cat more like the chunky three-dimensional cat I was drawing than it is like your two-dimensional paper picture of a dog? Yet it is supposed to ‘resemble’ the former and not the latter.

2) I can represent my cat but get things badly wrong (hence intuitively the representation is dissimilar). I might also happen purely by accident to hit upon a representation very similar to my neighbour’s dog. According to the theory I thereby represent that dog, despite my intentions.

3) I can have a mental representation of an abstract idea such as dog. But nothing can resemble all and only dogs. Images of dogs, for example, cannot have the generality of the abstract idea, and must indicate some properties that not all dogs have (an image of a dog is likely to be more like some wolves that it is like some dogs).

4) Schematic representations such as graphs can be successful with no need for similarity in any obvious sense to the represented states of affairs.

In short: similarity seems neither necessary nor sufficient for representation.
Functional Role Semantics

Go back to a question left hanging before. If propositional attitudes are not relations to propositions, how are propositions supposed to get into the story at all? Consider heights again. Treating them as if relations to numbers works because the order-structure of the family of height-properties is isomorphic to the order-structure of the real numbers—i.e. the structures can be mapped onto one another (the relations between height properties mirror the numbers). So if one property from the family is chosen for the unit height (e.g. 1 foot), then we can pick out another height property as that which stands to the unit height as the number six stands to the number one. We can in this way use numbers to index height properties according to their relation to this unit height.

Similarly for propositional attitudes: mirroring the network of causal relations there is a network of inferential relations between propositions (connecting propositions to those they logically imply or are implied by: e.g. connecting ‘Bill is taller than Ben’ to ‘Ben is smaller than Bill’, ‘Bill is taller than Ben or Bert is taller than Ben’, ‘Bill exists’, ‘At least one person is later than Ben’ and so on). The content of a given belief is indexed by the proposition which sits at the point in the inferential network of propositions mirroring the point at which that belief sits in the causal network of propositional attitudes. (But how plausible is the thought that there is this sort of mirroring?)

If this line can be made to fly, then the semantics (i.e. content or satisfaction conditions) of a propositional attitude is the proposition indexing the causal functional role that attitude has.

1. But if the content of a belief is determined by all the causal liaisons of my belief (holism) then it seems that when causal liaisons change my belief-contents change. Further, no two people will actually have share any belief (e.g. about the name of the Prime Minister). It seems there can be different overall causal roles, same contents. While if only a selection of causal liaisons is relevant to content-fixing, which ones and why?

2. On a natural reading, the account is radically internalist (semantics is fixed by what is going on in the head). But suppose I’ve met Jack Jones and you’ve met his identical twin James Jones, then our “Mr Jones” beliefs may have the same relevant functional role for each of us; what’s going on functionally in our heads is the same, yet my belief is about Jack and yours is about James. Same internal role, different contents.

Causal role plus external liaisons no doubt fix content, in this sense: two people with a set of intentional states internally related in exactly the same way, embedded in the environment in the same way, would have the same beliefs and desires. But semantics isn’t tied to all the fine details of causal role/external hook-ups.

Reliable Correlation Accounts of Representation

[Arguably originating in Locke (reacting against similarity accounts): more recently found in Dretske, Fodor and others.]

A mental state is a representation of something (or of some state) if it is regularly and naturally produced by that thing (or state). E.g. the state I am in when seeing white things represents whiteness because it is naturally caused by whiteness. Resemblance is not required: it is neither necessary nor sufficient for the correlation between the perceptual state and the worldly condition it represents.

Not all correlations are representations (e.g. exposure to sun and sunburn are correlated, but sunburn doesn’t represent the sun). So a reliability theory will have to filter our candidates for content (e.g. by their function in causing behaviour), with reliable correlation coming in to give the contents of candidates.

An initial version: state $x$ of a subject represents $y$ (e.g. redness) if and only if $x$ is produced when and only when the subject is confronted by $y$. But then there is no room for misrepresentation. The content of the representation turns out to be exactly those things causing it, so if at some time a green thing has the effect on me that red things usually have, then that thing also becomes part of what $y$ signifies.
Second shot: state $x$ of a subject represents $y$ if and only if in normal (or ideal) circumstances $x$ is produced when and only when the subject is confronted by $y$. Misrepresentation occurs when there is a malfunction, or when circumstances (e.g., lighting conditions) are abnormal. But …

1) Can we (without circularity) spell out what normal/ideal circumstances are?

2) If I learn the meaning of ‘tiger’ from pictures of tigers, does my concept pick out pictures of tigers rather than tigers themselves? And what about representations of unicorns? They can’t be states reliably correlated with unicorns because there aren’t any.

3) My tiger representations are also reliably correlated with certain patterns of stimulation on the retina. What account can be given of why the content is the tigers themselves rather than those patterns of stimulation? [The depth problem]

4) Will this capture all cases of misrepresentation? E.g., if I systematically misrepresent shrews as mice? My representation is more reliably correlated with mice-or-shrews than with mice. [The disjunction problem]

Fodor’s Asymmetric Dependence response to (4): a state represents mice because non-mice-caused ‘mice’ representations are asymmetrically dependent upon mice-caused ‘mice’ representations. There being non-mice-caused ‘mice’ representations depends on there being mice-caused ones but not the other way round. A dash of teleology helps with (1) and (3).

**Teleological Accounts**

Teleological claims focus on the *purpose* or function of a device. (‘Purpose’ doesn’t mean that there has to be a designer: natural selection suffices. The function of a device within a biological system is its effect which was responsible for its natural selection, and thereby its existence.)

The rough idea: A representation is of the thing or condition in the world which it has evolved to indicate. E.g., the orientation of bee dances results in spectator bees flying off in that direction. When all goes well they find flowers which leads (via pollen etc.) to their survival. ‘Flowers in that direction’ thus counts as what the oriented dance represents. This is compatible with misrepresentation on certain occasions.

A teleological account may be combined with a reliable correlation account to pick out those reliable correlations relevant to representation [e.g., Dretske; this could help with the depth problem]. Other writers [e.g., Millikan] allow reliable ‘if-and-only-if’ correlation to drop out of the picture. To ensure natural selection of a danger warning the representation should be caused when the danger is present, but it could be frequently caused at other times as well (a quick-and-dirty system erring on the side of safety could promote survival but thereby obstruct reliable correlation).

But what about e.g., for my beliefs about xylophones? In what sense have these evolved to indicate xylophones? Maybe there are simple ‘innate’ representations central to survival, like the bees dancings, but most of our mental representations are not of this type.

A bit more carefully (Papineau’s version, *Philosophical Naturalism*): A desire’s satisfaction condition is that effect which it is the desire’s biological purpose to produce. The purpose of the desire for sex is to get sex (though the desire may have other effects – make your pupils dilate, for example—these are side effects, not part of the biological purpose). Beliefs have purposes in a more indirect way. A belief’s purpose is to guide actions in such a way that desires are satisfied—i.e., so that desires fulfil their purposes. But a belief will only guide actions to success if it is true. So purpose goes with truth-conditions. In other words: a belief’s truth-condition is that condition which guarantees that actions generated by that belief will fulfil the biological purpose of satisfying desires. (So, it isn’t that beliefs about xylophones have been specially selected for: but we have a general mechanism for generating beliefs whose purpose is to successfully guide action.)

Imagine a duplicate of you coming suddenly into existence. Since its mental states are not due to a system acquired by natural selection, the teleological account seems to be committed to denying that they have representational content at all. But intuitively the duplicate would have beliefs and other mental representations too. (The teleological account is committed to an historical account of representation. How plausible is this as a global thesis?)

*Naturalism Again*
Note finally that although functional role, reliable indicator and teleological accounts probably won’t do as they stand, their failure is certainly no reason to suppose that a naturalistic account of intentionality isn’t possible, or that some composite theory won’t deliver the goods.

Reading

Abbreviations: see p. 2. Also

Introductory:
Churchland MC: Chapter 3.4.
Crane MM: Chapter 1.

Key reading:

An excellent overview is to be found in
M. Davies [PM]: 275–300.

Further reading: