17.C: the ***Philosophyldeas*** overview of

Functionalism

Behaviourism shifted the focus from the mind's inner nature, and explained it in terms of what it *does*. The consensus has been that this behavioural account is inadequate, because we need an *explanation* of the behaviour, in terms of the inner workings of the mind. The scientific spirit of the theory can be preserved, though, if we still focus on what the mind does, leaving neuroscience to track its internal structure. The model that suggests itself is computer software, which can be displayed in a flow diagram. We can ignore the fine-grained messiness of brains, and track the abstract relations of input, inner procedures, and output. The result can be a complete and understandable picture. Robots and animals have very different hardware from us, and yet seem able to implement similar mental events to humans (such as navigation, pain and memory). Hence mental activities are said to be '**multiply realisable**', and viewing the mind as software explains this, just as one program can run on a variety of computers. Thus functionalism has largely replaced behaviourism, offering a scientific account of the *inner* behaviour the mind – one which can avoid the need for dualism, or for mysterious inner qualities and forces.

The proposal, then, is that a mind consists of a set of interlocking functions, which produce behavioural outputs (including speech and body language), in response to complexes of inputs. The details of the brain or machine (or spirit) which implements the functions are irrelevant, since all that matters is what is 'realised' by the underlying system. So the mind is not a physical entity, but an **abstract** system (like being a 'vehicle', which has one function but is realised in many ways). The original proposal said the mind is a 'Turing Machine', which is the earliest theoretical model of a computer. It may even be that the mind has an inner language, resembling a computer's machine code, to run its procedures.

The system can be described at basic (perhaps physical), functional and mental **levels**. Each of these levels has its own properties, with our normal language about minds (called 'folk psychology', because it is deeply embedded in ordinary language) describing the most abstract top level. The properties of the functional level are all relational, and a mind is said to be constituted by its relations. This is not new, since faculties such as memory, desire and logic are only known by their role in thought. The question of how the basic level produces the functional level is not part of the theory, which just says that the mind is the functions. An objection was that each separate function of such a mind seems to be a mini-mind of its own (implying the need for a regress of explanations), but a response to that is '**homuncular**' functionalism, which says minds are indeed made of mini-mind components (each presumably with a will of its own), but with no regress, because they dwindle down to pure mechanisms at the simplest level. The mind is the emerging result of this team of focused functions.

Discontent with this highly abstract '**machine**' functionalism led to a version which was more firmly linked to the physical brain. Instead of treating mental relations as abstract, they were now treated as causal. In the analogy with computer software, the flow diagram of a program is a timeless abstraction, but on a computer this is turned into causal links, and the software makes things happen, as do real minds. Thus '**causal**' functionalism brought the theory closer to the human minds or brains that we want to explain. The theory takes the concepts of folk psychology and removes mysterious objects such as 'passions' or 'motives', replacing them with potentially observable patterns of cause and effect. Because the relations are causal, the functional level is identified more closely with the physical level, so that a pain just is the neurons associated with it. The pain is still multiply realisable, because other physical systems could play the same role, and the pain is not identical with the neurons, but with what those neurons do.

In functionalist theories, a mental event is said to be identical with a function, playing a particular causal role (however it happens to be implemented). Thus pains warn of danger, memories recall information, passions produce motivation. But how do you identify the role of mental events when the system is malfunctioning? There are phantom pains, false memories, and chaotic passions. A modified theory says a pain is a function which plays the appropriate pain-role in the system – it does what pains are supposed to do. This is '**teleological**' functionalism, which identifies a mind with a set of purposeful co-ordinated functions, having the joint purpose of supporting the survival and flourishing of some creature (or robot, perhaps). Typically this theory is integrated with the natural selection theory of evolution, which explains the implicit purposes of each feature of a mind.

For a time various forms of functionalism swept the board as the preferred theory of mind among philosophers. But the neat analogy with the procedures of computer software was not entirely convincing, because software and hardware are not distinct in computing (in a chip, for example), and because software output needs **interpretation** by a human user. One notable challenge said that if you set up an operator in a room which translated a language, simply by looking up in a catalogue the output for each input, the room might succeed in its task, but there would be no **understanding** in the room. Further doubts arose about the functionalist theory when other limitations of purely functional mechanisms began to be noticed. The theory implied that minds are just computers, but current computers don't understand, are not **creative**, lack **motivation**, have no **values**, are unaware of what they are doing, are not self-critical, don't attach **meaning** to their procedures, and struggle to interpret quite simple inputs. The central concern is that no one knows how a computer (or a functional system) might ever become **conscious**. Even if you implemented the flow diagram of a a human mind with electronics, we doubt whether that would produce consciousness. One response to this is to say that consciousness is unimportant (it's an 'epiphenomenon' – an output that doesn't do anything), but the common prejudice rejects this idea. If consciousness is optional, that would allow the implausible possibility of **zombies** (perfectly normal but non-conscious humans).

The possibility of an **inverted spectrum** (that if you and I look at violets and marigolds, your colour experience might be the reverse of mine, without us ever realising it) suggests that the causal role of colours could be preserved, and yet the experience of them be quite different, and functionalism says causal role is everything, so it could not explain that difference. Functionalism remains influential, but few now think it is the whole story.