Introduction to functionalism in the philosophy of mind

Graham Seth Moore

What is the relation between the mind and the body? Or, more specifically, what is the relation between mental states (the belief that it is sunny outside, the pain that you feel when you stub your toe) and states of the brain (the conditions of your neurons)?

It is often said that the answer lies within an analogy to computers. According to this popular idea, the mind is like a computer program that runs on the "hardware" of the brain. Just as computer programs are high-level patterns in the workings of a physical machine made out of metal and plastic, so too are mental states: they are patterns in the physical workings of the brain.

However, it must be admitted that, as it stands, this idea is highly vague. If this nascent idea is to be converted into a workable theory of the mind, then there are several things that must be clarified. For one, we must be clear about the actual relationship between computer software and hardware. If this relationship is supposed to be illuminating to the relation between mind and body, then we must say what it is. Secondly, we must say which features of the software-hardware relationship are shared by the mind-body relationship. Thirdly, we must spell out the consequences for our understanding of mental states.

Functionalism, in the philosophy of mind, represents a broad family of theories that attempt to do roughly this. There are several versions of functionalism. But, roughly speaking, each of them aims to spell out the relationship between the mind and the body by saying that mental states are abstract patterns that are implemented by physical states and processes. For us, these patterns are exhibited in our brains. But for the functionalist, a machine made from any material could, in theory, implement the patterns that are characteristic of mental states.

1 Instances vs kinds; psychoneural identity theory; multiple realization

It is important to begin by clarifying the question that functionalism is intended to answer. To this end, we must distinguish between instances of a mental state and mental state kinds.[1]

Consider *pain*. When we speak of pain in general, we are speaking of a kind of mental state. It is a kind of mental state because it can be shared by multiple people at multiple times. I stub my toe and I experience pain; you hit your funny bone and you experience pain. In each case, there is a common *kind* of mental state that we share: namely, pain.

Even though we share a common kind of mental state, each of us experiences our own separate *instance* of it. I have my instance of pain, which occurs at a particular time, and you have your instance of pain, which occurs at another particular time. So there are two instances of one kind of mental state.

Since there are two instances (my pain and your pain), we can ask: What do these two instances have in common? Why do they belong to the same mental category? In short: What do instances of pain have in common that makes them describable as *pain*?

This is the question that functionalism seeks to answer. It aims to identify the factors that unite the instances of each kind of mental state. This goes for pain, and it also goes for other mental state kinds: e.g. itching, experiencing the colour red, experiencing the colour blue, believing, desiring, etc.

Before we outline the functionalist answer, it would be helpful to first consider a different answer from one of functionalism's main opponents. By doing so, we can then explain the functionalist position by contrast.

The psychoneural identity theory is another theory in the philosophy of mind, and much like functionalism, it also attempts to say what the instances of a mental state kind have in common. The central tenet of the psychoneural identity theory is that *mental states just are brain states*. Mental states are *literally* identical to brain states, according to this view. This means, in particular, that mental state *kinds* are brain state *kinds*. Thus, according to the psychoneural identity theory, the instances of a kind of mental state must also be instances of a common brain state.

To illustrate this idea, philosophers often use the example of pain and c-fiber stimulation. C-fibers are a kind of nerve fibers, located in the brain, which are activated (or "stimulated") whenever a person experiences the sensation of pain. We know that when a person is in pain, their c-fibers are stimulated, and when their c-fibers are not stimulated, then they do not experience pain. (Anesthesia works by freezing the nerve pathways that lead to the c-fibers.) We thus say that c-fiber stimulation is *correlated* with pain, at least in humans. We know this through scientific observation.

The psychoneural identity theorist takes an additional philosophical step and hypothesizes that pain *just is* c-fiber stimulation. According to their view, pain and c-fiber stimulation are not two different kinds of states that happen to occur together (like smoke is correlated with fire); rather, they say that pain and c-fiber stimulation are literally *one and the same thing* (just as water is the same thing as H_20). This theorist would also make similar claims about every other kind of mental state. Each mental state is conjectured to be correlated

[1] In philosopher's jargon, this is the distinction between *tokens* and *types*.

with some kind of brain state, and the identity theorist claims that each mental state is identical to their correlated brain state.

Since they claim that mental state kinds are brain state kinds, the psychoneural identity theorist offers a straightforward answer as to what the instances of a mental state kind have in common: they are instances of a common brain state. So, what do all instances of pain have in common, according to this view? The identity theorist says: each instance of pain is a stimulation of c-fibers. If mental states are physical states, then it follows that each mental state is united by a common physical condition.

As it turns out, this last feature of the psychoneural identity theory also happens to be its greatest weakness. If pain is the same thing as c-fiber stimulation, then it follows that every creature that is capable of feeling pain must possess a brain that is endowed with this specific kind of neural fiber. However, it has seemed to many philosophers that this is overly restrictive. Afterall, there are many animals that have nervous systems that are physiologically quite different from ours, and yet they seem perfectly capable of experiencing pain. Take octopi for example. The fibers that make up an octopus's brain are physiologically distinct from the fibers that make up ours. But despite this, it seems absurd to deny that octopi feel pain.[2]

We can also consider hypothetical examples. Perhaps it is possible that aliens could have evolved with non-carbon-based bodies. Perhaps there could be aliens whose bodies are made of silicon. We can imagine these aliens having all of the same overt behaviour as humans. When they are pricked with a sharp object, they wince and groan and pull away from the cause of the damage. Moreover, we can further imagine that their silicon brains have mechanisms that trigger this avoidance behaviour whenever they suffer surface damage. In that case, these aliens wouldn't have *c-fiber stimulation* (since their "brains" are made of different materials), but it still seems quite plausible to say that they have *pains.*

Again, the problem with the psychoneural identity theory is that it implies that these aliens cannot feel pain, since their brains are made of different material from ours. But that just seems wrong. It seems entirely possible that a silicon-based brain could give rise to instances of pain, provided that it operates in a similar fashion to our carbon-based brains. The *material* (physical-chemical composition) that a creature is made of does not seem relevant to whether it can instantiate mental states. Rather, what matters to mentality is what the states of an organism can *do.* At least, that is the thought that motivates functionalism.

[2] Here is the argument against the psychoneural identity theory made explicit:

1 If pain is c-fiber stimulation, then only organisms with c-fibers will be capable of experiencing pain.

2 But some organisms—e.g. octopi—are capable of feeling pain and yet do not have c-fibers.

C Therefore, pain is not c-fiber stimulation.

This argument against the psychoneural identity theory also has the virtue of illustrating one of the core tenets of functionalism. This is the claim that mental states are *multiply-realizable.* By this, the functionalist means that the instances of any kind of mental state may be instantiated by many different kinds of physical states. To continue with our running example, the functionalist will say that pain can be "realized" by one kind of neural state in humans, a different kind of neural state in octopi, and yet a different kind of physical state in aliens. (A precise definition of "realized" will have to wait until the next section.) If I stub my toe right now, then the pain that I feel is a product of the c-fiber stimulation in my brain. And if an alien stubs its toe, then, according to the functionalist, the pain that it feels is the product of a different physical state: let's call it "A-fiber stimulation" (for "alien fibers"). So pains in humans are a matter of c-fiber stimulation, and pains in aliens are a matter of A-fiber stimulation. The point is, pain does not need to be identified with a single physical condition that is common to all pains, says the functionalist. Rather, there are many different physical conditions that can give rise to pain. Hence why they say that pain is *multiply* realizable.

2 Functionalism in general

So if, as functionalism says, pains do not need to have any particular brain state in common, then what else might they have in common that unites them together? Here is the second major tenet of the functionalist's picture. According to functionalism, mental states are defined by their *function*—i.e. by their *role* within the overall system of psychological processes and bodily behaviours of the organism. To put it succinctly, mental state kinds are conceived of as functional kinds. The instances of mental states are grouped by their common function.

Before we go on to sharpen this idea, it will be helpful to see an analogy. There are many concepts that we use in everyday life that are defined in terms of a function. Take, for example, the concept of a *mouse trap.* What makes something a mouse trap? What do all mouse traps have in common that makes them *mouse traps*?

Is it their physical material? No. Mouse traps can be made out of metal, wood, plastic, or virtually anything else. It would be wrong, therefore, to identify this kind of artifact with something defined solely by its physical and chemical anatomy. Mouse traps cannot be grouped together solely by their physical properties. There must be something else, besides their physical and chemical material, that makes something a mouse trap.

So what should we say instead? For the case of mouse traps, the answer should be obvious. The thing that all mouse traps have in common is what they *do* (not what they're made of): they are supposed to catch mice. That is to say, they perform a certain cause-and-effect pattern: they cause mice to become caught.

We can think of this in terms of input and output. The input is any situation in which a mouse is present. A mouse trap's job is to transform this situation into a state in which a mouse is caught—that's the output. Any device that can do that job of transforming a *mouse-is-present* state into a *mouse-is-caught* state will therefore count as a mouse trap. This can be done in many different ways. (Some traps are humane and allow the user to catch and release the mouse; others are not humane.) But provided that the device does this job in one way or another, it counts as a mouse trap.

Functionalists about the mind claim that similar things can be said about mental states. According to their view, each kind of mental state ought to be characterized by its function—that is, the role that it plays in transforming certain inputs into outputs.

To see what this means, consider the case of pain. In order to give a functionalist account of pain, the first step is to identify its typical causes and effects. In this case, the typical causes of pain will include: getting hit by a blunt object, getting poked by a sharp object, burning, joint inflammation, muscle tears, etc. On the other side of the equation, the typical effects of pain include: wincing, groaning, favouring the damaged or inflamed area, desiring to get rid of its cause, becoming distressed, believing that the area has been damaged, etc. Since we're only aiming to illustrate the general idea, it is okay if we leave these lists open-ended.

Once we have identified the causes and effects of pain, we are then in a position to state its functional role. According to functionalism, the function of pain is to be activated by its typical causes (tissue damage, etc.) and then to bring about its typical effects (wincing, groaning, desiring to be rid of it, etc.). This is, accordingly, what all pains have in common: they mediate between the transition from these causes to these effects. Just like a mousetrap is defined by what it does (bringing a *mouse-is-present* state into a *mouse-is-caught* state), pain, for the functionalist, is also defined by what it does (bringing a *tissue-damage (& etc.)* state into a *wincing-groaning-desiring-to-get-rid-of-it (& etc.)* state). To put it succinctly, what is it for a state to be a pain? It is for this state to be caused by tissue damage (etc.) and to bring about wincing, groaning, and other pain responses.

The functionalist will have similar things to say about every other mental state. If M is a mental state, then it has its typical causes $\langle \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3,... \rangle$ and its typical effects $\langle \mathrm{E}_1, \mathrm{E}_2, \mathrm{E}_3,... \rangle$. M can thus be said to have the function of facilitating the transition from $\langle \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3,... \rangle$ states to $\langle \mathrm{E}_1, \mathrm{E}_2, \mathrm{E}_3... \rangle$ states. The functionalist's core thesis, then, is that this function essentially characterizes M as a kind of mental state. What do the instances of M have in common? For the functionalist, the answer is that each instance of M facilitates the transition from $\langle \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3,... \rangle$ states to $\langle \mathrm{E}_1, \mathrm{E}_2, \mathrm{E}_3,... \rangle$ states. They are grouped together by the common function that they perform.

2.1 Realization

A function is an abstract pattern in the causal relations exhibited by a system, whether that be a person, an animal, or a computer. We have seen how the functionalist appeals to these abstract patterns to characterize the mental state kinds. This now puts us in a position to appreciate another aspect of functionalism. We can now ask: how do mental states relate to physical states?

We have already mentioned the name that the functionalist has given to this relation. They say that mental states are "realized" by physical states. Now that we have explained their view, we can finally say what "realization" amounts to.

Basically, for the functionalist, a kind of mental state is defined by its cause-and-effect pattern. For each particular organism (or non-organic system) that instantiates the mental state M, there will be some physical state P that facilitates this causal process. We can say that the physical state "plays the role" that defines M. This is what "realization" amounts to, according to the functionalist. When a physical state P exhibits the cause-and-effect pattern that defines mental state M, then P realizes M.

Once again, we have noted that for humans, c-fiber stimulation is correlated with pain. The functionalist will notice that c-fiber stimulation is not only correlated with pain, but it is also *caused* by the typical causes of pain and produces the typical *effects.* (E.g. it is caused by tissue-damage, and it causes wincing, groaning, etc.) For this reason, they will say that c-fiber stimulation *realizes* pain *in humans.*

This is consistent with pain being realized by other physical states in other systems. If we supposed that there are martians with silicon "brains" that contain a different type of circuitry—A-fibers—that are stimulated as a result of surface damage (etc.) and cause behaviour like wincing and groaning, then since they play the defining causal role of pain, it follows (by functionalist lights) that A-fiber stimulation realizes pain in these martians. And if some other physical state plays the causal role of pain in other organisms (e.g. octopi), then it will realize pain for those organisms.

This is how the functionalist explains the multiple realization thesis that was the downfall of the psychoneural identity theory. Earlier it was claimed that mental state kinds, like pain, can be exhibited by multiple different kinds of physical states and processes. According to the functionalist, this is because mental state kinds are defined by their function, and multiple different physical states can fulfill these functions.

3 Machine Functionalism

Functionalists define kinds of mental states in terms of patterns of causes and effects. Left as it is, this is fairly vague. But there are various ways to make it more precise. One such way is called *machine functionalism*, which was the earliest version of functionalism. This theory claims that mental states are exactly like the internal states of computers, or *Turing Machines.* In this section, we will spell out this claim in more detail.

The central concept of machine functionalism is that of a *Turing Machine.* (This concept was first defined by the British mathematician Alan Turing.)

Turing Machines are abstract descriptions that describe the functioning of a physical machine, but they abstract away from the physical details.

Before we define the concept of a Turing machine in full generality, it would be helpful to introduce the concept by giving an example of a relatively simple species of Turing machine. (This is called a finite state machine; it is just one kind of Turing machine among others.)

First of all, every Turing machine is a *discrete state machine.* This means that their operations can be described as a series of discrete states. At time 1 they are in one state, at time 2 they are in another state, and so on. Moreover, the state at any given time will be determined by various features of the state at the previous time.

Other than that, we can define the simplified kind of Turing machine by specifying (1) A (finite) set of possible inputs: I_1, I_2, I_3,..., (2) A (finite) set of internal states: S_1, S_2, S_3,..., and (3) A (finite) set of outputs: O_1, O_2, O_3,.... Once these are specified, we can then define the operations of the machine by specifying a function (in the mathematical sense) that maps input states and internal states to output states and internal states. In other words, it maps pairs of the form $\langle \mathrm{I}_i, \mathrm{S}_n \rangle$ to pairs of the form $\langle \mathrm{O}_j, \mathrm{S}_m \rangle$. This function will describe how the machine will behave over time—how it will transition from state to state, given its inputs.

The best way to get a sense of this idea is to give a concrete example. Let's imagine a very simple machine: a vending machine that dispenses coffee for a price of $1 per cup. To keep things as simple as possible, let's suppose that it only accepts change, and that it only accepts loonies ($1) and quarters ($0.25). In that case, there are two possible inputs:

I_1 A loonie is inserted.

I_2 A quarter is inserted.

There are five possible outputs:

O_1 Don't dispense coffee

O_2 Dispense coffee; give no change

O_3 Dispense coffee; give $0.25 change

O_4 Dispense coffee; give $0.50 change

O_5 Dispense coffee; give $0.75 change

Finally, to program this coffee machine, we will need to give it internal states, which can informally be thought of as:

S_1 Requires $1

S_2 Requires $0.75

S_3 Requires \$0.50

S_4 Requires \$0.25

With the input, output, and internal states specified, all that remains is to define the function that will describe the operations of the machine (how it moves from state to state). For example, if the machine is in S_1 (it has not yet received any change) and someone inserts a loonie (I_1) then it will dispense a coffee and give no change (O_2), and the machine will go back to its initial state (S_1) for the next customer to come along. If, on the other hand, the machine has not yet received any change (S_1), and someone inserts a quarter (I_2), then it will not dispense any coffee (O_1), and it will go into state S_2 (it will require another \$0.75). If the machine is in S_2 and someone inserts a loonie (I_1), then the machine will dispense a coffee and give \$0.25 in change (O_3), and it will go back to S. And so on. You can probably fill in the rest of the details of how this machine will work.

Now we can proceed to the general concept of a Turing machine. In general, Turing machines are much like what we have described above, except that they allow for a potentially infinite number of distinct inputs (so they are a lot more versatile) and they work by computation—that is, by manipulating symbols according to rigid set of rules.[3] Each Turing machine must have components that do the following jobs: (1) a "tape", which records the input, memory, and output, (2) a "head" (scanner / printer), which "sees" what is on the tape and prints symbols on the tape (depending on the symbol on the tape and the internal state) (3) a (finite) system of symbols ("alphabet") to print on the tape, and (4) a (finite) set of internal states. If a machine has parts that can effectively do each of these jobs, then it is capable of computation.

It is provable that every kind of computation can, in principle, be performed by a Turing machine. Thus the idea of a Turing machine captures the general idea of computation. If you think that the mind and brain work like a computer, then the idea of a Turing machine is crucial to making this precise. This gives us reason to be especially interested in Turing machines.

Turing machines are abstract mathematical functions that are defined by the relationship between their various components (inputs, outputs, internal states, symbols on their tape, the head, etc.). But what is the relation between these abstract functions and the physical machines that carry them out? Take, for instance, our coffee vending machine. What is the connection between the function that defines a coffee vending machine and an actual, concrete device made out of metal and plastic?

The answer is that the physical, mechanical processes of a concrete coffee machine *mirror* the abstract, mathematical relations that define a coffee vending machine in general. The physical causes and effects are *isomorphic* to the abstract pattern of inputs, internal states, and outputs.

[3] The input of a Turing machine is the initial state of its tape. The output is the final state of its tape.

A concrete coffee vending machine will be composed of various physical parts. Among these physical parts, it must have some part that "records" what internal state it is in (S_1, S_2, S_3, or S_4). If the coffee machine is made with modern digital technology, then it will record its state using electrically chargeable bits. But it could also record its internal state using more basic technology. Perhaps instead it could have a gear that turns on an axis and has four possible positions. It doesn't really matter what physical mechanism the coffee machine uses to record its internal state. The only thing that matters is that its other physical mechanisms are affected by its internal states in such a way so that the machine produces the right outputs (i.e. dispensing coffee and returning the right change).

This highlights a very important feature of machine functionalism. The internal states of a Turing machine are defined abstractly. They are defined by the function between inputs and internal states to outputs and internal states. This means that there are very few limitations on what kind of physical states can play the role of the internal states of a concrete embodiment of a Turing machine. (Indeed, from the standpoint of a computer programmer, the physical details of the internal states don't matter.) The *only* restriction on the physical realizers of the internal states is that they have the right causes and effects, which, overall, have the pattern described by the Turing machine.

Finally, now that we have the general idea of a Turing machine, we can return to the topic of the mind. As we've already alluded to, machine functionalists conceive of the mind along the lines of a computer. To put this more specifically, they identify *mental states* with the *internal states* of a Turing machine. Each person has a (huge!) series of "inputs" (sights, sounds, tactile impressions, chemicals ingested, and so on) and "outputs" (bodily movements, vocalizations, bodily reactions, and so on). Between these inputs and outputs, there is a huge set of complicated processes that run through the body and the brain. According to the machine functionalist, it is possible (in principle, although likely not in practice) to describe these processes in terms of a Turing machine. Once this is done, the machine functionalist will claim that our mental states *just are* the internal states posited by the Turing machine that describes all of our operations.

So then, what is *pain*, according to the machine functionalist? They will tell us that pain is a certain internal state in the Turing machine that describes the workings of all human beings and other pain-sensitive creatures. Pain is *realized* by physical states (e.g. c-fiber stimulation) by virtue of these physical states having cause-and-effect relationships that are isomorphic to the functional relations of this internal state (in this Turing machine).

4 The Ramsey-Lewis Method

I have mentioned that there are several different versions of functionalism in the philosophy of mind. Machine functionalism is only one version, but there are others. Machine functionalism is explicitly committed to the idea that mental processes can be captured by the concept of *computation* (as defined precisely by the concept of a Turing machine). But one can be a functionalist without endorsing that idea. One can instead say that mental states are defined by their causal roles generally, whether or not this is spelled out in terms of computation. Some functionalists place less emphasis on the analogy between minds and computers.

Regardless, there is another aspect of functionalism that needs to be highlighted. This is the idea that a mental state cannot be defined in terms of physical states *on its own.* Instead, according to functionalism, a mental state is defined by its characteristic function, and this includes *its relations to other mental states.* Thus the definition of a mental state must refer to other mental states.

Consider pain again. Earlier we remarked that its typical effects include winces and groans, which are bodily events. But pain also causes *the feeling of distress* and *the desire to get rid of it*, which are mental states. So to list the characteristic causes and effects of pain, we must mention other mental states in our definition. Functionalists generally agree that it is not possible to characterize the causes and effects of mental states purely in terms of bodily stimuli and bodily behaviour. (This is one respect in which they differ from another, older view: *philosophical behaviourism.*)

In that case, the functionalist definition of any given mental state will mention other mental states. And their definition of those other mental states will mention even more mental states. And so on. It would seem, therefore, that the functionalist never explains what a mental state is in non-mental terms.

This is a problem for anyone who is interested in *reducing* mental states to non-mental states, or characterizing mental states in terms of more fundamental phenomena, like brain states. It would seem to imply that we cannot understand mental states except by using circular definitions.

However, there is a method for overcoming this problem. It was developed by David Lewis and inspired by Frank Ramsey. Basically, the functionalist does not characterize each mental state in terms of a non-mental state *one-on-one*, on an individual basis. Instead, they give an entire *theory* of how *every* mental state is interrelated, and then reduce this entire *theory* to non-mental terms, in one fell swoop. This is a holistic approach to reducing mental states. It contrasts with the non-holistic approaches taken by psychoneural identity theory and philosophical behaviourism.

I will explain how this works by using a toy example. It would be impossible to show how this technique works in practice, because it requires us to give a theory of how *every* mental state is interrelated through their causes and effects. Obviously, such a theory would be enormous and complicated. So, instead, we'll consider a very short psychological theory that is unrealistically simple.

Imagine that this is our theory of the causes and effects of pain:

T For any person, **pain** is caused by damage to the skin, and it causes them to yell, become **distressed**, and **desire for the pain to go away**. **Distress** causes perspiration. **The desire for the pain to go away** causes one to move away from the cause of the damage.

(To reiterate, this is obviously a serious oversimplification of the causes and effects of pain.) Once we have made the theory explicit, we can then see the other mental states that are involved in our functionalist definition of pain (they are written in boldface). For this simple theory, we have mentioned *distress* and *the desire for the pain to go away.*

Now that the theory is made explicit, we can reduce the entire theory to non-mental terms by replacing each mental state term with a variable. Here is what our reduced theory will look like:

T* There are three internal states, $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$, such that, for any person, $\mathbf{M}_1$ is caused by damage to the skin, and $\mathbf{M}_1$ causes them to yell, to enter states $\mathbf{M}_2$ and $\mathbf{M}_3$. $\mathbf{M}_2$ causes perspiration. $\mathbf{M}_3$ causes one to move away from the cause of the damage.

There are two things to notice about this definition. First of all, notice that it no longer explicitly mentions any mental states. Sure, it involves variables that have replaced the terms for mental states, but it no longer mentions pain, distress, or the desire to be rid of pain. Secondly, notice that this definition still displays the causal roles of pain, distress and the desire to be rid of pain, even though it doesn't mention these states by name.

For the second reason, the reduced theory is particularly interesting to the functionalist. According to Lewis, it allows the functionalist to define each mental state in non-mental terms. Here is how to do it:

Pain = the internal state that occupies the role of $\mathbf{M}_1$ in T*.

Distress = the internal state that occupies the role of $\mathbf{M}_2$ in T*.

The desire for the pain to go away = the internal state that occupies the role of $\mathbf{M}_3$ in T*.

This, in a nutshell, is the Ramsey-Lewis method for giving functionalist definitions of mental states. The important takeaway is that each of the three mental states—*pain*, *distress*, *the desire to be rid of pain*—has been defined in terms of its causal role, just as functionalists claim they should be. Moreover, the final definitions offered do *not* explicitly mention any mental states, so they are suitable for explaining mental states in non-mental terms.

However—and this is a crucial thing to notice about this method—the price to pay for these definitions it that they are all interrelated through the theory T*. *Pain* cannot be defined on its own. To define *pain* in this way, we must also define *distress* and *the desire to be rid of pain.* And vice versa; to define *distress* in this way, we must also define *pain* and *the desire to be rid of pain.* To define any one mental state, we must also define them all. This is what it means for the definition to be holistic.

In summary, the Ramsey-Lewis method allows the functionalist to define mental states in terms of their causal role in a way that ultimately anchors the definitions to physical causes and effects. But to do this, we must define all of the mental states at once, by appealing to a total theory of *all* mental causes and effects. We either define all mental states altogether, or not at all.

5 Objections to functionalism

Functionalism (in some version or another) is perhaps the most widely endorsed solution to the mind-body problem by philosophers and cognitive scientists. Indeed, it is sometimes said to be the metaphysical foundation to cognitive science. For this reason, it is particularly important to study any problems it may have. For if there are any insurmountable objections, then this would imply that there's something deeply wrong with the current approach to the scientific study of the mind. In this section, we will look at three objections that have been raised against functionalism.

5.1 Functionalism's holistic approach to the mind

In the last section we highlighted the holistic nature of the functionalist's understanding of mental states. Each mental state is characterized by its causal relations to other mental states, and those other mental states, in turn, are characterized by *their* causal relations to other mental states, and so on. Thus every mental state will be implicated in the characterization of every other mental state. Any given mental state—for example, pain—can only be understood by its relation to the whole range of mental states instantiated by an organism.

This creates a fairly serious problem when we want to compare mental states across different kinds of organisms. For if mental states are characterized holistically, then it follows that two organisms can only share a kind of mental state *if* they share *every other kind of mental state*. But that is highly unrealistic!

Take pain for example. Intuitively, we think that pain is a mental state that can be shared by both humans and chickens. But we do not think that pain would have exactly the same causal role in humans and in chickens, all things considered. Pain in a human can cause them to desire to visit a doctor or take pain medication. But pain does not cause these states in chickens. In all likelihood, chickens are not even capable of having the desire to visit a doctor or the desire to take pain medication.

But according to functionalism, pain *just is* a state that's characterized by its causal role. So if humans and chickens aren't capable of exactly the same causes and effects (e.g. the desire to take pain medication), then it follows from the functionalist theory that they aren't capable of instantiating the same kind of mental state. Functionalism thus appears to entail that chickens (and other non-human organisms) aren't capable of pain after all.

There is a huge degree of irony in this objection. You may recall that functionalism was motivated, first and foremost, by the desire to explain the multiple realizability of mental states. The downfall of the psychoneural identity theory was supposed to be its inability to explain how both humans and octopi feel pain, and functionalism was supposed to save the day by explaining this fact. But according to this objection, it looks like functionalism can't explain multiple realizability after all. At least, not unless it makes some modifications.

5.2 The hivemind problem

There are about 86 billion neurons in the human brain. Given the rate of population expansion, it is conceivable that there will be a time, not long from now, when there will be this number of people. So imagine a world in which there are 86 billion people. And now imagine that they all work together in concert, and arrange themselves spatially in a way that's isomorphic to the arrangement of neurons in a brain. Finally, imagine that each person sends signals to their neighbors in a way that mimics the information being sent between the neurons in a brain.

If done properly, in perfect precision, it is possible that these 86 billion people could perfectly imitate the causal relations occurring between the neurons of a normal human brain. They would therefore mimic the causal relations that occur between stimuli, brain states, and bodily responses. If they were to do so, would it follow that they collectively instantiate the same mental states as humans? Would the collection of all of these actors constitute a *hivemind*?

The philosopher Ned Block has proposed this thought experiment as an objection to functionalism. His intuition is that the hivemind would *not* have the full range of human mental states. Specifically, it is hard to believe that the hivemind would feel pain.

Sure, there would be a group of actors that are recruited to imitate the functions of human c-fibers. And these actors would imitate the simulation of c-fiber firing in response to imitations of the causes of c-fiber stimulation, and they would subsequently cause other actors to act out the effects of c-fiber stimulation. So the causal pattern of c-fiber stimulation, and therefore pain, would be perfectly preserved in the collective behaviour of the hivemind. Therefore, there are states of the hivemind that play the causal role that is associated with pain. But does the hivemind feel pain?

Since, according to the functionalist, pain is essentially characterized by its causal role, it follows that, on their account, the hivemind feels pain. But to many philosophers, this seems like the wrong result. The actors in the hivemind are merely *imitating* the neural correlates, and causes and effects, of pain. But there is no single *person*, over and above the actors, who is experiencing pain. If this intuition is correct, then functionalism makes a wrong prediction.

5.3 The inverted spectrum objection

The final objection to functionalism pertains specifically to phenomenal mental states. These are the states that are intuitively characterized by *what it's like* to experience them. They have a particular *feeling* or *sensation* or *qualitative character* to them. Examples of phenomenal mental states include: *pain, pleasure, itching, seeing the colour red, seeing the colour green*, and many others.

The functionalist claims that these states (like all others) are characterized by their role in mediating between certain causes and effects. So what it is for a state to be an instance of *pain* is for it to be caused by bodily injury (etc.) and for it to cause typical pain responses.

But here's a problem. Ordinarily speaking, it would seem that the most essential feature of pain is not its causes and effects, but *how it feels.* The state of pain *feels painful.* It has a recognizable qualitative feeling to it, and that's what is essential to pain. If a state doesn't feel painful, then it simply isn't pain—even if it has the typical causes and effects of pain. However, this highly intuitive thought goes against the claims of functionalism.

The basic objection being raised here amounts to this. Each phenomenal mental state has both an associated feeling and an associated function. But it is possible for the feeling and the function to come apart. It is possible for there to be a state that has the feeling but not the function; and it is possible for there to be a state that has the function but not the feeling. In either case, we're inclined to think that each phenomenal state is tied together with its feeling, not its function. Therefore, phenomenal states are not characterized by their function.

One way to drive this point home is to consider the possibility of "inverted colour spectra." Have you ever wondered whether other people experience colours in exactly the same way as you do? When I look at something that is red, there is a particular experience of *redness* that is familiar to me. And no doubt, when you look at the same red object, you have a particular experience of redness that is familiar to you. We both call our experiences "the experience of redness". But perhaps your experience of redness and my experience of redness are not the same.

It's even possible to imagine that our internal experiences of colour are opposite of each other's. Imagine this possibility. When we both look at a red object (and we both call it red and say that we're having an experience of redness), the internal experience of the colour that I have is just like the internal experience that you have when you see something green. It's as if your 'red' is my 'green', and vice versa.

Is this scenario possible? Of course, we could never *know* whether there is such a thing as inverted colour spectra. The way in which we experience colours is private to each individual. I cannot see from my point of view what colours look like to you. And you cannot see what colours look like to me. Nonetheless, it seems possible that there could exist two people in reality whose experiences of colours are opposite to each other's.

If inverted spectra are possible, then this poses another problem for functionalism. The reason for this is that two people with inverted colour spectra will have different mental states—different colour experiences—*and yet their states will have exactly the same causes and effects.* Let's say that they both look at a red stop light. Then their visual systems will each cause them to go into their own respective state, with one colour experience for one person and a different colour experience for the other person. Both of their states will then cause exactly the same responses. After all, their behaviour may be indistinguishable (if everything else is equal). Both of them will *say* that they “saw a red stop light” and say that they “experienced the colour red” and they will both press the breaks and bring their cars to a stop.

Therefore, any two people with inverted colour spectra will have distinct mental states that nonetheless share the same causal role. But then it follows that their mental states cannot be grouped together solely by their causal role. If that’s the case, then functionalism, which seeks to group together mental states by their causal role, is on the wrong track. If inverted colour spectra are possible, then that proves that there’s more to phenomenal mental states besides their functional role within the cause-and-effect patterns in human psychology.