Introduction

Simone Gozzano and Christopher S. Hill

1

The psychophysical identity thesis asserts that psychological states are strictly identical with physical states of the brain. Versions of this view can be found in various figures in the history of philosophy, such as Lucretius and Thomas Hobbes, but it came to prominence in contemporary philosophy with three articles that appeared in the 1950s – Place (1956), Feigl (1958), and Smart (1959). More recently, it has been defended in Hill (1981), Hill (1984), Loar (1990), Hill (1991), Hill (1997), Papineau (2002), McLaughlin (2003), and Polger (2004). This theory was regarded as the standard solution to the mind–body problem in the fifties and early sixties. Then, a few years later, in the late sixties, it was summarily abandoned. Today, however, the psychophysical identity theory is being considered with a renewed interest. Many scholars are critically re-examining the arguments that have been marshalled against it, and are finding that it has the resources to strengthen a more comprehensively physicalistic metaphysics. There is also renewed interest in relations between mental and physical types as a result of developments in neuroscience and cognitive science. The chapters in the present volume continue this discussion. Some are concerned with questions about the proper formulation of the view; some seek to delimit its scope; some examine the motivation for accepting it; some explore strategies for defending it against objections; and some discuss its role in explanations and, more generally, in scientific practice. They all celebrate the virtues of the view, though some refrain from endorsing it, and one maintains that, in the end, its virtues are outweighed by its liabilities.

It is necessary to distinguish between two forms of the psychophysical identity thesis. One form maintains that concrete mental events, or *tokens* of mental states, are identical with concrete neural events, or *tokens* of brain states. On this view, it is true, for example, that the pain that Jones

is currently experiencing in her right arm is identical with a neural event, perhaps one in her somatosensory cortex. The other form claims that *properties* of concrete mental events, or mental *types*, are identical with physical *properties* of neural events, or neural *types*. This second form of the psychophysical identity thesis implies the first form, but it goes much farther. Thus, in addition to implying that Jones's current pain is identical with a concrete event in her brain, it implies that the property *being a pain* is identical with a neural property, perhaps the property *being a certain sort of somatosensory activity*. It is generally held that this second view, which is known as *type physicalism* and also as *type materialism*, is much more interesting, and correspondingly more deserving of philosophical study, than the first, which goes by the name of *token physicalism*. Thus, since it is much stronger than token physicalism, its ability to unify, simplify, and systematize our theories of mind and human nature is much greater. Not surprisingly, it is the only form of the psychophysical identity thesis that receives positive attention in the present volume.

Still, even though token physicalism is less appealing than type physicalism, it might seem that it is a live option, and that it should be held in reserve in case its more ambitious cousin should fail. This comparatively sanguine view is challenged in Jaegwon Kim's contribution to the present volume. Kim argues that any virtues token physicalism might be thought to have are illusory.

Most advocates of type physicalism have seen it as concerned exclusively with mental states that have a qualitative dimension. To be more specific, they have maintained that qualitative properties like *being a pain* and *being a sour taste* are identical with neural properties, and have denied that this is true of intentional properties like *being a belief about Cleopatra* and *being a visual experience of Cleopatra*. Roughly, the reason for the denial is that intentional states generally have contents that involve extra-cranial objects and properties. For example, to believe that Cleopatra was of Macedonian descent is to be in a state that involves relations to Cleopatra and to ancient Macedonia. On the other hand, it is often thought that qualitative properties are purely internal. But which properties count as qualitative? It is difficult to answer this question in any sort of final way, but it will serve present purposes adequately to describe them as properties that, according to folk psychology, cannot be grasped fully unless one has been directly acquainted with them – that is, unless one has experienced them from a first-person perspective. They include properties of bodily sensations, properties of perceptual experiences, properties of emotional experiences, and properties of images.

Most of the contributions to the present volume are concerned exclusively with properties of these kinds.

But this is not true of all of the contributions. In recent years, philosophers of neuroscience have become interested in the question of whether the processes and mechanisms that are studied by cognitive science can appropriately be identified with neural processes and neural mechanisms. The processes and mechanisms with which they are concerned typically lack a qualitative dimension. Thus, for example, one of the questions that has recently received a lot of attention is whether the consolidation of memories can be identified with certain types of neural activity. (Consolidation is the process by which passing events become established as enduring records.) It is clear that consolidation must eventuate in structural changes in the brain, but is it *identical* with the neural process that produces the relevant changes? Or does it stand in some lesser relationship to the process, such as supervenience or realization? Further, if it is in fact appropriate to view consolidation as identical with a neural process, on what level is the relevant process to be found? At the level of large-scale neural networks? At the level of individual neurons? At the molecular level? Questions of this sort have generated a provocative and rapidly expanding literature. Several of them are explored in the present volume, particularly in the chapters by William Bechtel, John Bickle, and Robert McCauley.

In order to simplify the exposition, we will in the following sections use the expression "psychological properties" in a restricted sense – specifically, to stand for the psychological properties with which advocates of type physicalism are currently concerned. Thus, psychological properties will include qualia and the properties of processes and mechanisms that philosophers of neuroscience take to be strongly reducible to neural properties. They will not include beliefs, desires, and other intentional states.

1 1

There are four reasons to prefer type physicalism to alternative views. First, as J. J. C. Smart emphasized in his early papers on the topic, it is simpler than alternatives, because it sees only one category of properties where other theories see two. Here is Smart's well-known formulation of this point:

> If it be agreed that there are no cogent arguments which force us into accepting dualism, and if the brain-process theory and dualism are equally consistent with

the facts, then the principles of parsimony and simplicity seem to me to decide overwhelmingly in favor of the brain-process theory. (Smart 1959, p. 156)

Second, it has more explanatory power than the various forms of dualism. Unlike dualism, it can reductively explain the large array of laws in which psychological properties are involved, including the correlation laws that link psychological properties to neural properties. Third, it does a better job of honoring our intuitions about the causal powers of psychological states than do other theories. We believe, for example, that pains are causally responsible for much of our thought and talk about pain, and also for such forms of behavior as wincing, crying out, and taking steps to secure relief. Type physicalism sustains all of those intuitions, and does so in an especially simple and straightforward way. Fourth, it is implied by a body of knowledge that consists of a priori principles about the causal roles of psychological properties and a posteriori claims about the causal roles of neural properties. We will say a bit more about each of these four considerations.

The simplicity argument invokes Occam's Razor, which advises that entities are not to be multiplied beyond necessity. This principle is widely thought to provide a rationale for preferring type physicalism to property dualism. Even dualists are inclined to agree that if one theory is more complex than another, then its advocates bear the burden of proof. It remains to be seen, however, whether Occam's Razor provides an *epistemic* reason for accepting type physicalism or a reason of some other kind. One might think it obvious that the Razor provides an epistemic reason. After all, the difference between dualism and type physicalism is just that the former goes beyond the latter in its existential commitments, making all of the existential claims that type physicalism makes and one more as well. Or so it can seem. On this view of the matter, it appears that dualism makes a stronger claim about reality than type physicalism, and that one therefore takes more of a risk in believing it. (Since dualism makes a stronger claim, it is less likely to be true.)

Reflection shows, however, that these observations neglect an important dimension of the relationship between the two theories. It is true that dualism claims that there is an irreducible category over and above the irreducible categories posited by type physicalism, but it does not follow from this that type physicalism makes a weaker claim than dualism. Dualism and type physicalism are alike in asserting that reality contains a category consisting of psychological properties and also a category consisting of neural properties. If dualism seems to be a more ambitious theory than type physicalism, this is because, after making this claim about

categories, dualism goes on to assert that the two categories are mutually irreducible. But type physicalism goes on to make an additional claim of its own – specifically, that one of the categories *is* reducible to the other. It is not at all clear that a claim of reducibility is weaker than a claim of irreducibility, and by the same token, it is not at all clear that one would take less of a risk in accepting physicalism. Accordingly, it may be a mistake to see the simplicity argument for physicalism as fundamentally epistemic in character. Perhaps it should be seen as an aesthetic argument instead. In explaining his commitment to simplicity, Quine once said that he had a taste for desert landscapes. This can't be all that there is to the matter, for the appeal of simplicity is more universally appreciated than the beauty of the desert. But it may be necessary to think of simplicity, in the relevant form, at least, as more closely related to beauty than to probability or truth. (See Hill 1991, pp. 29–40.)

As noted, the second argument for type physicalism emphasizes the explanatory power of the doctrine. This argument has two versions. The first version begins with the assumption that there are strong correlations between psychological states and certain neural states. (Accordingly, it presupposes that the multiple realization argument, which is discussed in the next section, can be answered.) It then claims that type physicalism provides the best explanation for these correlations. Thus, for example, it claims that the best way of explaining the correlation between pain and a certain brain state is to say that pain is identical with that state. Finally, it invokes the best explanation principle, which asserts, roughly speaking, that one is entitled to believe a theory of *X* if the theory provides the best explanation of all of the data that are relevant to *X*. The conclusion is of course that we are entitled to believe type physicalism. (See Hill 1991, pp. 22–26, and McLaughlin 2010.) The second version of the argument is like the first; but instead of invoking correlations between qualitative states and neural states, it invokes laws linking qualitative states to other phenomena, such as behaviors of various kinds. Consider, for example, the generalization that pain causes one to withdraw reflexively from aversive stimuli. It would be nice to be able to explain this generalization. According to the second version of the argument, we can provide such an explanation if we suppose that pain is identical with the brain state that is the neural cause of reflexive withdrawals. Indeed, it is claimed, we can provide the best explanation of the generalization if we make this supposition. But if this is so, then the best explanation principle authorizes us to accept the supposition. (See Block and Stalnaker 1999.)

There has been considerable interest in both versions of the explanatory power argument in recent years. Here we will just reply briefly to an objection to the first version that has often appeared in the literature. According to the objection we have in mind, it is a mistake to say that correlations can be explained by saying that the correlated items are identical. The idea here is that if *X* is the very same thing as *Y*, it is a logical error to say that *X* and *Y* are correlated. There cannot be a correlation unless the correlated items are distinct. This objection rests on a misunderstanding. The first version of the explanatory power argument is concerned with correlation *laws* – that is, with *propositions* of the form "An instance of *X* occurs when and only when an instance of *Y* occurs." There is no doubt that propositions of this form can be fully meaningful, and fully true, even if it should turn out that the property to which "*X*" refers is identical with the property to which "*Y*" refers. Now a philosopher of mind is confronted with the question of whether it is possible to derive certain propositions of the given form from more fundamental propositions. It appears that the answer to the question will be "yes" if there are grounds for accepting the corresponding propositions of the form "*X* is the very same thing as *Y*." The latter propositions imply the former. Also, they are more fundamental than the former propositions, because, if true at all, they are necessarily true, and are therefore not in need of explanation. That is, they are more fundamental because they can bring a chain of explanations to an end. (Perhaps it is worth observing in this connection that identity propositions can have the status of laws of nature. This is true, for example, of Newton's second law, and of Einstein's observation that $e = mc^2$. Presumably claims like "Pain is identical with brain state *B*" can share this status.)

Ansgar Beckermann and Brian McLaughlin continue the discussion of the explanatory power argument in their contributions. The chapter by Christopher Hill is also concerned with the correlations between qualitative states and brain states, but it focuses on the question of how much room they allow. Do they force us to view qualia as properties of internal states, or can they somehow be accommodated by theorists who prefer to see sensory qualia as properties of bodily states, and perceptual qualia as properties of external objects?

The third argument for type physicalism is based on the perception that mental causation is robustly real – more specifically, the perception that psychological states play essential roles in the causal histories of various forms of behavior, and also in the causal histories of other psychological states. It seems that any sound metaphysical theory should sustain this intuition.

But it is not clear that theories other than type physicalism are capable of sustaining it. Consider the generalization that pain causes one to cringe. Because we have general inductive grounds for believing that physical phenomena always have physical causes, and because cringing is a physical phenomenon, we know that cringing is caused by a certain brain state – say, *B*. Now if pain is identical with *B*, there will be nothing mysterious about the fact that both pain and *B* cause one to cringe. Saying that pain causes one to cringe and that *B* causes one to cringe will just be two different ways of saying the same thing. On the other hand, if pain is distinct from *B*, then we will have to say that pain's causal efficacy with respect to cringing merely duplicates that of *B*. But how can the causal contribution of pain be essential if it merely duplicates the contribution of *B*? Moreover, if pain is distinct from *B*, it seems that we will have to say that it brings cringing about without a continuous intervening process. That is, there will have to be some point *P* in the physical process running from *B* to cringing at which pain acts directly, without benefit of there being an intermediate process linking it to *P*. Accordingly, its causal power will be mysterious, like that of telekinesis. And there will be other problems.

This version of the argument from mental causation was originally put forward by Kim (1998). Kim's argument has received a great deal of favorable attention in the literature, but it has also been criticized on a number of grounds. Thus, for example, it has been maintained that it rests on an outmoded conception of causation. Alyssa Ney's contribution defends the argument, maintaining, among other things, that the presupposed notion of causation can play an important role in scientifically informed metaphysics.

The fourth argument derives from Lewis (1966). It can be summarized as follows:

Premise 1: Pain = the state, whatever it may be, that occupies the causal role *R*.

Premise 2: Brain state *B* = the state that in fact occupies causal role *R*.

Conclusion: Pain = brain state *B*.

Here "causal role *R*" stands for a collection of causal properties that includes *caused by tissue damage*, *causing distress*, and *causing withdrawal from an aversive stimulus*. Lewis maintained that the first premise is known to be true a priori. In effect, his idea was that "pain" is used to abbreviate a complex description. The second premise is shown to be true by empirical investigation, and the conclusion is inferred from the premises

in accordance with the principle that identity is transitive. Frank Jackson develops the argument in his chapter, putting flesh on the skeletal version that Lewis devised, and also defends it from a range of objections.

III

Although it is very appealing for the reasons we have been reviewing, type physicalism is challenged by a formidable array of objections. The multiple realization argument is one of the most impressive of these, and it has probably had the broadest influence. It was originally formulated and elaborated in several papers that Hilary Putnam published in the 1960s. (See, e.g., Putnam 1967.) Putnam's reasoning can be summarized as follows:

Premise 1: Where *P* is any psychological kind, there is a wide variety of creatures that can possess *P*, including members of other species and complex androids like C3PO.

Premise 2: If there is a wide variety of creatures that can possess *P*, then there is no one physical kind by which *P* is realized – at best, it is realized by different kinds in different creatures.

Premise 3: If *P* is realized by different physical kinds in different creatures, then *P* cannot be identical with any specific physical kind.

Conclusion: No psychological kind is identical with any physical kind.

By way of illustration, Putnam maintained that it is very unlikely that pain is realized by any one physical kind, because pain is common to animals, reptiles, and mollusks ("octopuses are mollusca, and certainly feel pain"), and the brains of these creatures differ radically in point of physical structure. Of course, if there is no one kind that realizes pain, it is true a fortiori that there is no one kind with which pain can be identified.

Putnam's argument is rightly viewed as one of the most significant contributions to the philosophy of mind of the twentieth century. Among other virtues, it provides the main motivation for functionalism, which maintains that psychological states are individuated by their causal roles, not by the structural or compositional features of brain states. It seems likely that most philosophers of mind think that functionalism provides the correct account of a broad range of psychological states. Still, impressive as it is, there is reason to think that the multiple realization argument has been overrated. We will briefly note a few objections.

First, Putnam seems to have overstated the case for the first premise. He says that members of a number of radically different species *certainly* feel

pain, and he seems to have held similar views about the distribution of other mental states, such as feelings of thirst and hunger. What led him to these conclusions? We know that he was not influenced by similarities between human brains and reptile brains, or by similarities between human brains and mollusk brains, for his second premise implies that any such similarities are less important than the differences between brains of these types. So he must have been relying on behavioral similarities. But we know today that behavioral similarities can be an untrustworthy guide to psychological similarities. Consider visually guided action. In normal human beings it seems to depend to at least some degree on conscious visual experiences. Thus, for example, one's ability to determine which way a pencil is tilting depends on conscious experiences, and this also true, to an even higher degree, of one's ability to negotiate complex landscapes. But we know today that blindsight patients are able to recognize tilt, and that victims of visual form agnosia are capable of very complex endeavors, such as hiking over difficult terrain. The fact that someone can navigate a room without bumping into furniture, or reach out and shake your hand, is not an adequate basis for drawing conclusions about the contents of the person's visual consciousness.

Second, it's possible to explain away the appeal of the first premise. It appears that there is a heuristic for attributing mental states that leads us to make provisional attributions on the basis of simple movements. To see this, recall the classic video made by the social psychologists Fritz Heider and Mary-Ann Simmel (Heider and Simmel 1943). In this film, two dark triangles and a dark circle engage in various behaviors that are naturally interpreted as aggression, pursuit, flight, observation, hiding, and bonding. Initially, at least, viewers find it almost impossible to refrain from interpreting the movements of the figures in terms of psychologically pregnant descriptions of this sort, and they tend to posit purely psychological underlying causes, such as affection, hostility, covetousness, fear, the desire to control, and the desire to escape. But all that happens on the screen is that abstract geometrical shapes move in various suggestive ways. Why is it so natural to interpret simple movements in terms of these complex mentalistic concepts? The obvious answer is that there is a more or less hard-wired heuristic for attributing mental states that takes account only of motions. This explains why viewers are drawn to mentalistic interpretations of the figures. Moreover, the thought that the attributions are due to a heuristic is deflationary, implying that they are provisional and subject to correction as information increases. And in fact, it seems that observers are inclined to withdraw the attributions if it is stipulated that the figures are simple two-dimensional shapes, without

an internal organization of any kind. Evidently, the initial attributions are hostage to discovery of an appropriate internal complexity. There seems to be a general pattern here. Thus, for example, we are willing to attribute pain to ants, but we tend to withdraw these attributions when we come to appreciate that the ants have none of the neural structures that support experiences of pain in human beings. (See Hill 1991, pp. 220–25.)

Third, as we noticed earlier, the multiple realization argument seems to commit us to a functionalist account of mental states; but there are grounds for doubting that qualitative states like pain can be identified with causal roles, for it is very difficult to find a set of causal powers that is present in all and only those cases in which a given qualitative state is present. Consider pain. Paralytics can experience pain, as can babies, masochists, and those with the disorder known as pain asymbolia. (Patients with this disorder insist that they continue to feel pain, but they maintain that their pains no longer bother them. Their testimony is confirmed by imaging studies, which point to lesions in the centers that are known to be responsible for the emotional dimension of pain experience.) Paralytics cannot engage in pain behavior; babies cannot form desires or beliefs about their pains; masochists differ from the rest of us in that they actively seek painful experiences; and asymbolia patients see pains as on a par with uninteresting tingles – they neither mind them nor find them especially worthy of attention. (For discussion, see Grahek 2001.) In view of these facts, it seems unlikely that there is a set of causal powers that is both necessary and sufficient for the existence of pains (Hill 1991, pp. 73–76).

Fourth, if the multiple realization argument appears sound, this may be because we are using mismatched principles of individuation for psychological kinds and realizing neural kinds. In a well-known discussion of this view, William Bechtel and Jennifer Mundale summarize it as follows:

> [O]ne diagnosis of what has made the multiple realizability claim as plausible as it has been is that researchers have employed different grains of analysis in identifying psychological states and brain states, using a coarse grain to identify psychological states and a fine grain to differentiate brain states. Having invoked different grains, it is relatively easy to make a case for multiple realization. But if the grain size is kept constant, then the claim that psychological states are in fact multiply realized looks far less plausible. One can adopt either a coarse or a fine grain, but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic. (Bechtel and Mundale 1999, p. 202)

Bechtel and Mundale support this claim by arguing that some of the kinds recognized by neuroscience are extremely broad, encompassing