The Mind-Body Problem in 20th Century Philosophy

Amy Kind
Claremont McKenna College
akind@cmc.edu

What is the nature of the mind? What is its relation to the body? These questions – which jointly constitute *the mind-body problem* – lie at the heart of philosophy of mind. Traditionally, there have been two sorts of approaches to this problem. According to the position known as *dualism*, the mind is an immaterial thing not existing in physical space. Dualists believe that there are two fundamental kinds of things in the world: material things, like trees and tables and chairs, and like our physical bodies, and immaterial things, like minds. The mind, according to the dualist, has a different kind of nature from the body (which includes the brain). In contrast, the position known as *monism* holds that there is only one type of fundamental entity in the world. The most influential form of monism, traditionally known as *materialism* but now more commonly referred to as *physicalism*, claims that all entities – including the mind – are physical in nature.

In the wake of René Descartes' influential 17th century arguments in favor of dualism, it was long assumed that physicalism was not a tenable position. Such was the general philosophical consensus about the mind-body problem at the close of the 19th century.[1] But this was to change in the 20th century, a period of time in which considerable attention was addressed to the mind-body problem, and also in which considerable progress was made. This essay, which aims to take a historical look at this progress, traces the progression of philosophical thought about the mind-body problem over the course of the 20th century.

Because it would be impossible in an article of this sort to survey all of the important developments on the mind-body problem that occurred in the 20th century, I here focus my attention on several of the key movements and themes that occupied philosophical attention over the course of the last hundred years. The first three sections trace the development of physicalism about the mind from behaviorism to the identity theory to functionalism. In the fourth section, I turn to the qualia-based threat to such theories that arose in the last quarter of the century.

1. Behaviorism

The story of the mind-body problem in 20th century philosophy begins with behaviorism, a movement that dominated philosophical thinking about the mind for at least the first half of the century. Insofar as behaviorism offered a genuine alternative to both traditional dualism and traditional

[1] See, e.g., the assessment by U.T. Place: "[E]ver since the debate between Hobbes and Descartes ended in apparent victory for the latter, it was taken more or less for granted that whatever answer to the mind-body problem is true, materialism must be false." (Place 2002, 36)

Forthcoming in the *Philosophy of Mind in the Twentieth and Twenty-first Centuries* edited by Amy Kind.

materialism, both of which had been found wanting, its popularity is perhaps unsurprising. Though behaviorist theories come in several different varieties, they all in some way attempt to understand the mind in terms of bodily behavior. While philosophers prior to the onset of behaviorism had long recognized a tight connection between mind and behavior, this connection had generally been understood to be evidential in nature. My reaching for a drink counts as evidence that I'm thirsty; my moaning and groaning counts as evidence that I'm in pain. In contrast, behaviorists argued that we should view the connection between mental states and bodily behavior not as evidential but as constitutive. For the behaviorist, we should not think of bodily behavior as a mere manifestation of some inner mental state; rather, exhibiting such behavior is simply what it is to be in the relevant mental state.

The behavioristic turn in philosophy at the start of the 20th century mirrored a similar turn in psychology. Dissatisfied with the introspectionist methods that had previously been dominant, psychological behaviorists aimed to reorient psychological study towards more objective methods that would put the discipline on a similar footing with other sciences. According to psychological behaviorists, psychology is best understood not as a science of mind but as a science of behavior. Though our primary interest here concerns behaviorism in philosophy of mind, it will be useful to begin with a brief discussion of psychological behaviorism. Doing so will help us to better understand the philosophical varieties of behaviorism.

1.1. Psychological Behaviorism

The term "behaviorism" was coined in 1913 by John Watson in "Psychology as the Behaviorist Sees It," an article often referred to as the behaviorist "manifesto." Though there had been some isolated expressions of a behavioristic bent among nineteenth century psychologists, psychology in the late nineteenth century and the early 20th century was largely a study of inner mental life.[2] Following the work of Wilhelm Wundt, often regarded as the father of modern psychology, this study was conducted by way of rigorous introspective investigation. As William James wrote in his *Principles of Psychology*, "Introspective observation is what we have to rely on first and foremost and always." (James 1890/1981, 185)

In urging that psychology should study behavior rather than the inner causes of behavior, psychological behaviorists were largely concerned with issues of scientific methodology. But many of them were also often tempted by a stronger stance, one that denies the existence of such inner causes altogether. This eliminativist tendency was particularly marked in the late-20th century work of B.F. Skinner, who referred to his view as *radical behaviorism* and contrasted it with the methodological behaviorism of psychologists like Watson. In *About Behaviorism*, where he dismissed alleged inner causes of behavior as "mental fictions," Skinner explicitly identified human thought with human behavior: "Thinking has the dimensions of behavior, not of a fancied inner process which finds expression in behavior." (Skinner, 1974, 18, 117-8). Ultimately, however, Skinner's support for eliminativism was not entirely unequivocal. For example, though he claimed in *Science and Behavior*

[2] See Titchener 1914 for a discussion of 19th century antecedents to behaviorism in psychology.

that statements like "he eats" and "he is hungry" both refer to the same behavioral fact, he also noted that the "objection to inner states is not that they do not exist, but that they are not relevant in a functional analysis." (42) As we turn to philosophical behaviorism, we will see a similar flirtation with eliminative behaviorism – and one that is similarly ambiguous.

1.2. Philosophical Behaviorism

While behaviorism was widespread among philosophers in the first half of the 20th century, there were really two distinct versions of the view on offer, each springing from a different motivation. The behaviorism associated with philosophers such as Rudolf Carnap and Carl Hempel was an outgrowth of logical positivism and the verificationist theory of meaning. Based on the supposition that there are close logical connections between statements involving mental vocabulary and statements involving behavioristic vocabulary, this view is typically referred to as *logical behaviorism*. The behaviorism associated with philosophers such as Gilbert Ryle and Ludwig Wittgenstein was an outgrowth of ordinary language philosophy. While this view does not have a standard name in the philosophical literature, I will call it *ordinary language behaviorism*.[3]

Logical behaviorism is a theory about the meaning of statements involving mental expressions – statements like "Diego has a toothache" or "Sofia believes that it will rain." While the meaning of these psychological statements may seem to depend on their reference to inner mental states – to Diego's toothache and to Sofia's belief – the logical behaviorists disagree. Instead, they take the meaning of such statements to consist in behavioral facts about Diego and Sofia, i.e., facts about the behavior that these individuals manifest or that they are disposed to manifest.

Underlying this view is a commitment to the verificationist theory of meaning, a theory that takes the meaning of a statement to be established by the conditions of its verification (see Hempel 1980, 17). Consider a claim about temperature, e.g., the claim that the current temperature in my office is 72 degrees Fahrenheit. For such a statement to be true, it would have to be the case that the mercury level of a properly calibrated glass thermometer currently placed in my office would correspond to the number 72 on a Fahrenheit scale. We could also make analogous claims about an alcohol thermometer or an infrared thermometer, or about various other devices; as Hempel notes, there is a long list of other possibilities that make the statement true. Each of these possibilities can be expressed by what he calls a *physical test sentence*. We need not establish the truth of all of the physical test sentences in evaluating the truth of the original sentence. But the key point is that the original sentence about temperature communicates to us nothing more than the fact that these physical test sentences obtain; the original sentence is simply an "abbreviated formulation" of such sentences. (Hempel, 17)

In identifying the meaning of a statement with the conditions of its verification, the verificationist is in turn committed to the claim that statements lacking verification conditions lack meaning. Though such a statement might be grammatically well-constructed, it lacks any content and is thus only a pseudo-statement. (see, e.g., Hempel, 17; Carnap, 44) What then of psychological

[3] In discussions of behaviorism, philosophers often adopt different classificatory schemes. See, e.g., Byrne 1994 and Graham 2010.

statements? Since there is no way in principle to test for inner states like pains and beliefs, must such statements be dismissed as meaningless? To avoid this result, the logical behaviorist suggests that psychological statements have verification conditions that are directly analogous to those we saw in the temperature example. Psychological statements are verified by facts about behavior. For example, the verification conditions for the claim that Diego has a toothache include physical test sentences like the following:

- Diego grimaces and rubs his mouth
- When asked, "What's wrong," Diego utters the words, "I have a toothache."
- Diego has swollen gums and a tooth with an exposed pulp

and so on. For the logical behaviorist, then, mentalistic vocabulary should not be taken to refer to inner mental states. Rather, the meaning of claims involving such vocabulary consists in facts about behavior.

In contrast to logical behaviorism, ordinary language behaviorism was not motivated by verificationism. Rather, the behaviorism of philosophers like Ryle and Wittgenstein was primarily grounded in worries about the problem of other minds, a problem that is particularly acute if there are inner mental states that are private to each individual. As Wittgenstein suggested in his posthumously published *Philosophical Investigations*:

> The essential thing about private experience is really not that each person possesses his own exemplar, but that nobody knows whether other people also have *this* or something else. The assumption would thus be possible – though unverifiable – that one section of mankind had one sensation of red and another section another. (Wittgenstein 1953, §272)

As he went on to suggest in the famous "beetle-in-a-box" passage, problems arise from the assumption that people understand a mental state like pain only from their own case:

> Suppose everyone had a box with something in it: we call it a "beetle." No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle.—Here it would be quite possible for everyone to have something different in his box." (Wittgenstein 1953, §293)

While there is considerable dispute about how best to interpret this passage (as well as the larger argument of which it is a part), we can nonetheless here see Wittgenstein's worries about how we would know anything about other minds if mentalistic vocabulary were to refer to private mental states.

Ryle expressed related worries in *The Concept of Mind,* a book that offered an extended attack on the view that the mind is an immaterial substance distinct from the body. On this Cartesian picture, one that Ryle often referred to derisively as "the Cartesian myth" or as "the dogma of the ghost in the machine," solipsistic worries naturally arise: "I can witness what your body does, but I cannot witness what your mind does, and my pretensions to infer from what your body does to what your mind does all collapse, since the premisses for such inferences are either inadequate or unknowable." (Ryle 1949, 60) To overcome such worries, Ryle urged that we see mental vocabulary as functioning to refer to behavioral dispositions: "To find that most people have minds ... is simply to find that they are able and prone to do certain sorts of things." (Ryle 1949, 61) Likewise, Wittgenstein too argued that once we pay careful attention to the way language is used, we see that it is a mistake to see the grammatical function of mental vocabulary as one of reference to mental states; verbal expressions involving the word "pain," for example, are simply instances of pain-behavior, no different from other instances of pain-behavior like crying. (See Wittgenstein1953, §244)

In developing their views, both Wittgenstein and Ryle at times seemed to embrace eliminativism. In the beetle-in-the-box passage, for example, Wittgenstein went on to note that "the thing in the box has no place in the language-game at all; not even as a *something*, for the box might even be empty." (Wittgenstein 1953, §293) Likewise, in dismissing Cartesianism as a myth – in claiming that the postulation of mind as an entity distinct from the body is a "category mistake" – Ryle also seems to be expressing sympathy for an eliminativist view. Ultimately, however, neither of these philosophers came down squarely on the eliminativist side. Wittgenstein explicitly pulled back from eliminativism when he noted that the respect in which mental states are fictions is that they are *grammatical* fictions; a sensation "is not a *something*, but not a *nothing* either!" (Wittgenstein 1953, §304) Similarly, though Ryle's scorn for talk of mentality and minds is apparent, his discussion tended to fall short of showing how, exactly, we can successfully analyze such talk away.[4]

1.3. Criticisms of Behaviorism

Despite the dominance of behaviorism in the first half of the century, however, in the 1950s and 1960s it came under attack from several different directions. A sharply negative review of Skinner's 1957 book *Verbal Behavior* by Noam Chomsky (1959) called psychological behaviorism into question. According to Chomsky, language acquisition and verbal competence cannot be explained simply in terms of stimulus and reinforcement; rather, we must postulate innate mechanisms to achieve an adequate explanation. Around the same time, important criticisms directed at both logical behaviorism and ordinary language behaviorism began to surface in the philosophical literature.

One influential criticism derives from the work of Roderick Chisholm (1957). For the behaviorist, belief consists in behavioral dispositions; for example, we might analyze a gardener's belief that it will rain in terms of his disposition to carry an umbrella with him while he works and to put away his watering can. (See Ryle 1949, 174) But such behavioral dispositions implicitly presuppose the presence of relevant desires: A gardener who believes that it will rain will carry an umbrella only if he wants to stay dry. Thus, any attempt to provide a behavioral definition of belief would have to make reference to desire and, likewise, any attempt to provide a behavioral definition of desire would have to make reference to belief. That such mental notions are inherently connected – that they form an "intentional circle" – dooms any attempt to define them solely in terms of behavior.[5]

[4] In the second half of the 20th century philosophy, non-behaviorist versions of eliminativism were developed by various philosophers. This kind of view, which seems to have its roots in the work of Wilfred Sellars, is notably found in W.V.O. Quine (1960), Paul Feyerabend (1963), Richard Rorty (1965), and Paul Churchland (1981).

[5] Chisholm 1957, 173-185. See also Geach 1971, esp. 7-9.

A second influential criticism derives from the work of Hilary Putnam, and in particular to the article, "Brains and Behavior." This criticism is aimed specifically at the logical behaviorists. To make the case that the kinds of analyses they offered were in principle unworkable, Putnam asks us to imagine a community of stoic individuals in which all of the adult members have trained themselves to entirely suppress their involuntary pain behavior. These super-spartans might occasionally verbally admit they are in pain – in a normal, pleasant tone of voice – but they will show no other sign. When they stub their toes or burn their fingers, they don't wince or moan, or flush or break out in a sweat, or grab the affected body part. Yet they still feel pain as we do, and they dislike it. Taking this one step further, Putnam next asks us to imagine a community of super-super-spartans. Having been super-spartans for so long, they no longer even make verbal reports of pain, and they will not admit to being in pain if they are asked. Because we can conceive of this sort of case – a case of pain without any pain-behavior whatsoever and, in fact, without even any disposition to pain-behavior – logical behaviorism must be mistaken.

The criticisms of behaviorism struck many as decisive, and by the late-1960s, behaviorism had largely disappeared from view. Though there are behavioristic elements present in the work of some late 20th century philosophers – perhaps most notably in the work of Daniel Dennett – the vast majority of contemporary philosophers reject the reduction of mind to behavior.[6] As we will see, however, behaviorism left an important legacy, for the rise of both the identity theory and functionalism in the second half of the 20th century can be traced in large part to the lessons learned in discussions of behaviorism.

2. The Identity Theory

Even while behaviorism was dominating philosophy of mind in the early part of the 20th century, both the philosophical and the psychological literature contained isolated expressions of a different sort of physicalist view, one that identifies mental states not with behavioral dispositions but instead with physical states of the brain. The philosopher Moritz Schlick, for example, claimed that we should not understand the relationship between our experience and brain processes as one of causality but rather one of simple identity. (Schlick 1925/1974) Likewise, the psychologist Edwin G. Boring claimed that "consciousness is a physiological event." (Boring 1933, 14) But it was not until the late 1950s that the identity theory achieved philosophical prominence. The rise of the theory owes almost entirely to the publication of three ground-breaking articles: "Is Consciousness a Brain Process," by U.T. Place (1956), "The 'Mental' and the 'Physical'," by Herbert Feigl (1958), and "Sensations and Brain Processes," by J.J.C. Smart (1959). As summarized by Feigl, the identity theory consists in the claim that "the states of direct experience which conscious human beings 'live through,' and those which we confidently ascribe to some of the higher animals, are identical with certain (presumably configurational) aspects of the neural processes in those organisms." (Feigl 1958, 446)

Feigl's development of the theory – which occurred while he was working at the University of Minnesota – proceeded separately from Place and Smart's development of the theory – which occurred while there were both at the University of Adelaide. There are thus various minor differences between

[6] See Dennett 1987 for essays in which his behavioristic tendencies are in evidence.

what's sometimes called the *American identity theory* and what's sometimes called the *Australian identity theory*. But these differences won't matter for our purposes here.[7] As Place himself notes, "Although there are certain differences of detail in the positions adopted in these three papers, the area of agreement was sufficiently great for all three of the original protagonists to be able to agree that they were all defending the same basic position." (Place, n.d.)

2.1. The Case for the Identity Theory

Prior to the 20th century, materialists had often identified mental states with various physical states. In antiquity, Democritus understood the soul as a sort of fire, made out of spherical atoms, and he took thought to consist in the physical movement of atoms. In the 17th century, Hobbes claimed that sensations are simply internal motions of the sense organs. But the identity theory of the 20th century departs from these previous theories in at least two key ways. First, in focusing on brain processes, the identity theorists aligned themselves with neuroscience. Given the tremendous advances in neuroscientific research in the 20th century, this alignment gave credibility to their theory. Second, and perhaps more importantly, the identity theorists took the psycho-physical identities they posited to be directly analogous to other scientific discoveries. Just as scientists discovered that lightning is identical to a certain kind of electrical discharge or that heat is identical to molecular motion, so too the identification of specific mental states with specific brain states emerges as a scientific discovery. We might discover, for example, that pain is identical to the stimulation of C-fibers.

It is worth pausing a moment over this particular example. Though it is now in widespread use in philosophical discussion, the three papers by Place, Feigl, and Smart that ushered in discussion of the identity theory did not invoke this particular identity claim, nor did they use other specific examples of this sort. Rather, they tended to talk more generally of a sensation being identical to some brain process or other. Reference to the pain/c-fiber identification did not become common in philosophical discussions of the identity theory until the 1960s (see, e.g. Putnam 1960 and Rorty 1965). Importantly, however, the use of "the stimulation of c-fibers" or "c-fiber firing" in such discussions seems best understood as a placeholder term, i.e., as a stand-in for whatever brain process is discovered to be identical with pain (assuming that any is). It should thus not be seen as a threat to the identity theory if it turns out that pain is not c-fiber firing but is some other kind of brain process (see, e.g., Puccetti 1977); the precise identity will be determined by scientific discovery.

In drawing an analogy to scientifically discovered identities, the identity theorists emphasized several related features of such identities, some epistemic and some semantic. For example, the identity theorists stressed that the plausibility of their theory hinges on the recognition that not all identities have the same epistemic status. Many identities – like the claims that "red is a color" and "a square is an equilateral rectangle" – can be known *a priori*. In contrast, an identity like "lightning is electrical discharge" can be known only *a posteriori*; it is an empirical claim that results from scientific inquiry. Psychophysical identities, said the identity theorists, are to be understood analogously to claims like "lightning is electrical discharge" rather than to claims like "a square is an equilateral rectangle."

[7] See, e.g., Crawford 2013 for discussion of the differences.

Claims like "pain is c-fiber firing" are also the result of scientific inquiry and thus cannot be known *a priori*.[8]

In making their semantic points, the identity theorists called upon early 20th century research in philosophy of language. Frege's seminal work on the distinction between sense and reference showed that two expressions that refer to the same object may nonetheless differ in meaning by having different senses. Consider the expressions "the Morning Star" and "the Evening Star." Though they both refer to the same object – the planet Venus – the sense of the former expression is something like "the first heavenly body visible in the morning sky," while the sense of the latter expression is something like "the last heavenly body visible in the evening sky." According to the identity theorists, the distinction between sense and reference comes into play in the case of empirically discovered identities. Though the word "lightning" refers to the same phenomenon as the words "electrical discharge," these two expressions do not have the same sense. Likewise, though the word "pain" refers to the same phenomenon as the words "C-fiber stimulation," these two expressions do not have the same sense. Thus, it's no objection to the identity theory that someone might be perfectly able to discuss his own pains and sensations without knowing anything at all about neuroscience or even about the brain – as Smart noted, "a person may well know that something is an A without knowing that it is a B" even though A is identical to B. Thus: "An illiterate peasant might well be able to talk about his sensations without knowing about his brain processes, just as he can talk about lightning though he knows nothing of electricity." (Smart 1959, 147)[9]

More generally, the identity theorists persuasively showed that many potential objections to the identity theory stem from similar confusions about the nature of psychophysical identities. Here it will be worthwhile for us to explore one other specific implementation of this general strategy, since this will also help to illuminate why the identity theory is often referred to as *the topic-neutral theory*. Consider after-images and, more specifically, the fact that we typically refer to them as being colored. For example, after staring at a bright green colored patch, a viewer who turns her attention to a white surface might plausibly describe her experience by saying something like, "I have a magenta afterimage." Call this claim *M*. Claims like M seem to pose a problem for the identity theory: Though the after-image is magenta, the correlated brain process is not, so how can the after-image be identical to the brain-process? In response to this worry, the identity theorist argues that, once claims like M are properly understood, they can be seen to be consistent with physicalism. Though the general line of argumentation owes to Place, the point was more forcefully developed by Smart. As he argued, rather than taking M to commit us to the existence of something magenta-colored, we should best understand it as having the (rough) meaning: *There is something going on that is like what goes on when (for example) I see a corncockle flower.* Understood this way, M contains only "quasi-logical" or "topic-neutral" words and does not presuppose that the after-image is immaterial. (Smart 1959, 150) In this way, the identity theorists argued that much of our purportedly mentalistic vocabulary – and indeed,

[8] The early identity theorists also claimed that such scientific discoveries were contingent. In the wake of Saul Kripke's *Naming and Necessity* (1980), this claim is now largely thought to be mistaken.

[9] Place (2002, 37) takes this feature of the identity theory to be central in distinguishing it from earlier versions of materialism.

our very experience of our own mentality – is actually non-committal between dualism and physicalism.[10] This insight proves critical to establishing the viability of the identity theory, and more generally, the viability of physicalism.

As our discussion thus far suggests, the initial case put forth for the identity theory was in many ways a defensive one. Feigl, Place, and Smart were typically more concerned to answer or forestall objections than to mount positive arguments for their view. Smart, for example, noted that the object of his paper was "to show that there are no philosophical arguments which compel us to be dualists." (Smart 1959, 143) Insofar as these early identity theorists put forth a positive case for the theory, it rested largely on considerations of Ockham's razor: Given theories of equal explanatory power, we have reason to adopt the one that is ontologically more parsimonious. As Smart put the point: "If it be agreed that there are no cogent philosophical 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."[11]

Seeds of a further positive argument lie in a worry originally expressed by Feigl. Immaterial mental states, were they to exist, would have to be "nomological danglers" (Feigl 1958, 428), i.e., they would remain entirely outside the system of physical laws. Later identity theorists further developed this argument, relying heavily on the thesis that physics is thought to be *causally closed*, or *complete*, i.e., the causal history of any physical event can be wholly given in physical terms. (See, e.g., Papineau 2002) This thesis, which seems immensely plausible in light of the scientific advances of the 19th and 20th centuries, deprives opponents of the identity theory of a plausible account of mental causation.[12] Intuitively speaking, mental causes play a crucial role in the causal histories of human actions: my desire for a drink causes me to get up from where I'm sitting and walk to the kitchen, my fear causes me to back up when I encounter a rattlesnake on the hiking trail, my toothache causes me to make an appointment with the dentist. If we accept the completeness of physics, however, then someone who denies the identity theory can account for mental causation only by accepting one of the following two unapalatable alternatives.

> (1) Human actions are always overdetermined, wholly and completely caused by mental events and also wholly and completely caused by physical events. Thus, even if I didn't have a desire for a drink, I would still have taken the same action.
>
> (2) The appearance of mental causation is an illusion. In reality, mental events are epiphenomenal, i.e., they have no causal power.

In contrast, the identity theorist's account of mental causation is perfectly in line with the completeness of physics. Since the identity theorists claim that mental events are identical to physical events, they

[10] For a useful discussion of topic-neutrality, see Armstrong 1979, 75-79.

[11] Smart does not work out this argument in any detail, but see Christopher Hill (1991, ch. 2) for a more comprehensive attempt to show that considerations of simplicity favor the identity theory over its dualistic rivals.

[12] See Chapter 7 for a detailed discussion of the problem of mental causation.

can explain human action in terms of mental causes without denying that a physical event's causal history can be given wholly in physical terms.

Generally speaking, then, the positive case for the identity theory can be seen as one of inference to the best explanation. According to the identity theorists, the best way to account for mental causation is to see mental causes as themselves physical. More generally, the best way to account for all of the undeniable psychophysical correlations that we observe is in terms of identity. There are not two distinct things whose correlation needs explanation; rather there is only one thing. As we saw earlier, a similar strategy of inference to the best explanation was employed by the behaviorists. But the identity theorists have a plausible reason to claim that the explanation they offer is better than the one offered by the behaviorists. In reducing mental states to behavior, the behaviorists had to deny that mental-state talk serves a reporting function. For the behaviorist, my claim that I am in pain does not serve to report my pain but rather is part of what constitutes it; consider Wittgenstein's remark that "The verbal expression of pain replaces crying and does not describe it." (Wittgenstein 1953, §244) As the identity theorists pointed out, this seems implausible. But unlike the dualist, who views such a claim as a report of an "irreducibly psychical something" (Smart 1959, 142), the identity theory can view the claim as a report that refers to a brain process (albeit perhaps unknowingly to the one who makes the report).

Though the identity theory in this way makes a considerable advance over behaviorism, the early identity theorists did not fully abandon the behaviorist leanings of the early 20th century. As originally developed, the identity theory was meant to apply only to experiential mental states, states like mental images and pains. With respect to other mental states like beliefs and desires, the early identity theorists thought that behaviorist analyses were largely correct. Place, for example, noted explicitly that for cognitive and volitional concepts "there can be little doubt ... that an analysis in terms of dispositions to behave is fundamentally sound." (Place 1956, 44) Later identity theorists like David Armstrong and David Lewis explicitly rejected the restriction of the theory to experiential states. Putting emphasis on the virtue of theoretical economy, these later theorists thought that it would be preferable to give a unified account of all mental phenomena (see, e.g., Armstrong 1968, 80). In further developing the identity theory, Armstrong and Lewis also emphasized the causal nature of mental states, thereby paving the way for the functionalist theories of mind that became prominent in the late 1960s and that continue to be prominent today.

2.2. Multiple Realizability

To understand the rise of functionalism, however, we must first understand an influential criticism directed against the identity theory in the late 1960s, namely, what we might call the *multiple realizability argument*. Forcefully developed by Hilary Putnam and Jerry Fodor among others, the argument rests on the claim that creatures with very different neural mechanisms might all feel pain, i.e., that pain might be multiply realizable in many different kinds of physical structures. As Putnam puts the point, the truth of the identity theory requires the existence of some type of state such that any creature whatsoever who is in pain is in that physical state. But creatures as diverse as mammals, reptiles, and molluscs all seem unquestionably to experience pain, despite having very different neural structures. And can't we conceive of extraterrestrial life forms who also experience pain? (Putnam 1967, 436) Here we might consider Lewis's example of a hydraulically powered Martian. Martian pain feels just like human pain, though the Martian has a physical constitution quite different from that of humans:

> His hydraulic mind contains nothing like our neurons. Rather, there are varying amounts of fluid in many inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others. His mental plumbing pervades most of his body—in fact, all but the heat exchanger inside his head. When you pinch his skin you cause no firing of C-fibers—he has none—but, rather, you cause the inflation of smallish cavities in his feet. When these cavities are inflated, he is in pain. And the effects of his pain are fitting: his thought and activity are disrupted, he groans and writhes, he is strongly motivated to stop you from pinching him and to see to it that you never do again. (Lewis 1980, 216)

In brief, the identity theory postulates an identity between *types* of states, with each type of mental state identified with a type of brain state. (For this reason, the theory is often referred to as *type physicalism*). My pain and your pain are both tokens of the type *pain*, but so too are the pain of an octopus and the pain of a Martian. For the identity theory to be true, all tokens of the type pain must also be tokens of the same type of physical state (be it the state of C-firing firing or some other state). But just as two token mousetraps (or two token clocks, or two token engines) might be made of – or *realized* by – very different physical materials, so too it seems that two tokens of the type pain might be realized by very different physical states.

The multiple realizability argument came to be seen as a significant threat to the identity theory. Importantly, however, that is not to say that the identity theory has been discarded. Unlike behaviorism, the identity theory continued to attract support through the final decades of the 20th century, and versions of the theory continue to be defended in these early days of the 21st century. Among the various strategies available for responding to the multiple realizability argument, one promising line retreats to species-specific reduction and concedes that human pain is a distinct type of mental state from, e.g., octopus pain. (For discussion, see Kim 1992.)

At this point, it's also worth noting that the threat posed to the identity theory by the multiple realizability argument is not a threat to physicalism in general. Nothing in the argument shows that pain is nonphysical, i.e., it is compatible with the argument that all token pains are realized by some physical state or other, even if those physical states are not all of the same type. Thus, for all we've said so far, *token physicalism* remains a viable theory.[13] Other objections that have been raised to the identity theory, particularly those concerning qualia, the phenomenal aspects of our mental states, do count against physicalism more broadly. But since such objections are best understood against the backdrop of both the identity theory and functionalism, we will postpone discussion of them until the final section of this paper.

[13] Donald Davidson's anomanlous monism is one particularly prominent version of token physicalism. See, e.g., Davidson 1970.

3. Functionalism

Behaviorism was threatened by the possibility that an organism might be in a mental state without exhibiting any of the characteristic behavior associated with that mental state, i.e., an organism might be in pain without exhibiting any pain behavior. The identity theory was threatened by the possibility that an organism might be in a mental state without being in the characteristic brain state associated with that mental state, i.e., an organism might be in the state of pain without being in the state of c-fiber firing. Functionalism manages to block both of these threats by treating mental states as functional states. For the functionalist, a mental state like pain is identified by the functional role that it plays in the life of the organism. While historical antecedents to functionalism can be found in the work of Aristotle and Hobbes, the view received its first detailed development in the second half of the 20th century.[14] Its rise coincides with important developments in computer science and particularly in artificial intelligence, and functionalists have often drawn on computational analogies in spelling out their position. On the functionalist view, mentality is better thought of at the level of software than at the level of hardware.

3.1. Mental States as Functional States

Above we noted that a device like a mousetrap can be realized in multiple different physical structures. For something to be a mousetrap, what matters is not what it is made of but what it does, i.e., the function it performs. The notion *mousetrap* must thus be specified not physically but functionally. In this way mousetraps are different from, say, nuggets of gold. For something to be a gold nugget, it must have a specific physical constitution, i.e., it must be composed of atoms with atomic number 79 – hence the truth of the expression, “All that glitters is not gold.” Compare pyrite, or fool’s gold, which has a similarly brilliant yellow luster but is a compound of iron sulfide.

Mousetraps are not the only things that are specified functionally. A similar point applies to many other artifact concepts – engines, clocks, pencil sharpeners – and even biological concepts. As Jaegwon Kim notes, “What makes an organ a heart is the fact that it pumps blood. The human heart may be physically very unlike hearts in, say, reptiles or birds, but they all count as hearts because of the job they do in the organisms in which they are found, not on account of their similarity in shape, size, or material constitution.” (Kim 2011, 131)

The functionalist claims that mental states are better understood on the model of the mousetrap than on the model of gold nuggets. Consider again the mental state pain. This state plays a certain role in the life an organism. It typically comes about because of bodily damage, and it typically results in wincing, moaning, avoidance behavior, fear, a desire for relief, and so on. Or consider the mental state thirst. It typically comes about because of lack of adequate hydration, and it typically results in dry mouth, liquid-seeking behavior, a desire for liquids, and so on.

As this suggests, the functionalist’s characterization of mental states is strikingly reminiscent of the behaviorist characterizations of mental states. In particular, both the functionalist and the behaviorist define mental states in terms of a relation between inputs and outputs. But despite this

[14] See Levin 2013 for a discussion of historical antecedents to functionalism.

similarity, there are nonetheless several important differences between the two kinds of characterizations. Unlike the behaviorist, the functionalist does not deny that mental states are internal states of the organism. For the functionalist, a statement like "I am in pain" does not count as just pain behavior along the lines of wincing and moaning but serves as a genuine report. This leads to a related difference between functionalism and behaviorism. By accepting that mental states are internal states of organisms, the functionalist can make reference to such states in the specification of inputs and outputs. Pain produces not only certain characteristic behaviors but also certain mental states; as indicated above, it typically leads to a desire for relief.

Our discussion thus far highlights two important tenets of the functionalist view. First, mental states are interdefined. Second, mental states are multiply realizable. The first point protects functionalism from many of the objections that threatened behaviorism; the second point protects functionalism from many of the objections that threatened the identity theory. While these two tenets underlie functionalism in general, the view comes in several varieties that differ from one another in various important respects.

As originally articulated by Putnam, functionalism was formulated in terms of a Turing machine, a hypothetical device proposed in 1936 by mathematician Alan Turing. (For this reason, Putnam's version of functionalism is often referred to as *machine functionalism*.) In brief, the operations of a Turing machine can be wholly characterized by a set of instructions given in what's often called a *machine table*. For each internal state of the computer, the instructions specify the output that will result from a given input. An example drawn from Fodor (1981) helps to elucidate the concept.[15] Consider a simple gumball machine that sells gumballs for a dime, takes both nickels and dimes, and is capable of dispensing change. The operations of the machine can be wholly described by the following table:

	Dime input	Nickel Input
S1	Dispenses a gumball and remains in S1	Proceeds to S2
S2	Dispenses a gumball and a nickel and proceeds to S1	Dispenses a gumball and proceeds to S1

As this table indicates, the machine has two possible states. Metaphorically speaking, we can think of S1 as the state *waiting for a dime* and S2 as the state *waiting for a nickel*.[16] The machine is waiting for a dime when it has received no money since last dispensing a gumball; the machine is waiting for a nickel when it has received a nickel since last dispensing a gumball. If the machine is waiting for a dime and it gets a dime, then it dispenses a gumball and continues to wait for a dime. If the machine is waiting for a dime and it gets a nickel, then it switches to waiting for a nickel. If the machine is waiting for a nickel and it gets a dime, then it dispenses a gumball and a nickel and switches to waiting for a dime. If the

[15] I have amended this example slightly.

[16] In describing the gumball machine's states this way, I do not mean to suggest that the gumball machine should be thought of as having mental states. To reemphasize, the description is meant to be metaphorical.

machine is waiting for a nickel and it gets a nickel, then it dispenses a gumball and switches to waiting for a dime.

For the machine functionalist, the mind can be thought of as a Turing machine, i.e., the operations of the mind can be completely described by way of a machine table. Each mental state corresponds to one line – perhaps a very long line – in the machine table. Though coming up with the appropriate machine table will undoubtedly be quite difficult, Putnam notes that the project of doing so – that is, the project of coming up with "'mechanical' models of organisms" – is an "inevitable part of the program of psychology." (Putnam 1967, 435)

Returning to the machine table above, note that while it gives a complete specification of the operation of the gumball machine it says nothing about its physical constitution. The gumball machine might be made of plastic, of metal, of wood, and so on. In fact, it might even be made of non-physical stuff. Consider Fodor's claim that: "As far as functionalism is concerned a [gumball] machine with states S1 and S2 could be made of ectoplasm, if there is such stuff and if its states have the right causal properties." (Fodor 1981, 129) Machine functionalists like Putnam tended to think the same could be true of the mind and hence took their view to be compatible with dualism (Putnam 1967, 436).

As functionalism has developed, however, it has tended to be classified as a physicalist view, and reasonably so: Most functionalists see themselves as committed to physicalism. The commitment underlying the physicalist version of functionalism might be captured as follows: While mental states may be realized in many different physical substances, they must all be realized in some physical substance or other. As this suggests, however, the physicalist version of functionalism – and hereafter it should be assumed that I am talking about this version of the view unless I explicitly note otherwise – is not a version of *type* physicalism. Rather, it is a version of *token* physicalism. For the functionalist, every token pain is realized in some physical state, but those physical states might be tokens of different physical types – perhaps c-fiber firing in humans while something altogether different in a hydraulic Martian.

In the wake of Putnam's work, various versions of functionalism have been developed in the philosophical literature. These subsequent versions retain the core commitment of functionalism – that mental states should be understood as functional states – while dropping the commitment to understanding functional states in terms of machine states. Some functionalists endorse *psychofunctionalism*, the view that mental states are defined by the functional roles they play in an empirical theory, specifically, that of cognitive psychology (see, e.g., Fodor 1968). Other functionalists endorse *analytic* or *conceptual functionalism*, the view that mental states are defined by the functional roles they play in our ordinary or 'folk' theory (see, e.g., Lewis 1966, Armstrong 1968). This version of functionalism emerges from logical behaviorism and shares its underlying motivation of providing analyses of our ordinary mental state concepts. Yet other functionalists endorse *teleological* *functionalism*. What's distinctive to teleological functionalism is the claim that that the notion of 'function' must be understood teleologically, i.e., in terms of biological purpose.[17]

3.2. Criticisms of Functionalism

By the 1970s, functionalism had become the dominant view in philosophy of mind, and in fact, even now at the beginning of the 21th century it continues to enjoy widespread acceptance. But despite its popularity, the view has nonetheless faced significant criticisms. One key strand of attack, emerging from the work of John Searle, claims that functionalism is unable adequately to capture the intentional nature of our mental states. In this context, intentionality doesn't have to do with intention but with *aboutness* or *directedness*. Consider my belief that Albert Pujols is a baseball player. This belief, which is about Albert Pujols, has intentional content. When I hope that Pujols will hit a lot of home runs next season, or when I desire his autograph, these mental states too have intentional content – they too are directed at Albert Pujols. Importantly, mental states can have intentional content even if they are directed at things that do not exist. Someone who has mistaken Conan Doyle's stories for nonfiction might admire Sherlock Holmes and desire his autograph. Though Sherlock Holmes does not exist, these mental states are intentional nonetheless.[18]

Searle's famous Chinese Room thought experiment aims to show that computers cannot achieve understanding and, correspondingly, that functionalism cannot provide an adequate account of mentality. Consider a computer that is programmed to speak Chinese. If the program were good enough – if, say, the program were to put the computer in the same functional states as a native speaker of Chinese – then the computer would produce outputs that were indistinguishable from such a speaker. The computer would appear to understand Chinese. But, says Searle, this appearance would be mistaken, for the mere instantiation of a program cannot endow a computer with understanding. To defend this point, Searle imagines that he is inside a room with a very sophisticated rulebook equivalent to the computer's program. When Searle enters the room, he does not understand Chinese, and has no idea what the different Chinese characters mean – they look to him like mere squiggles. The instructions in the rule book tell him what squiggles to output upon receiving certain other squiggles as input. But now suppose he gets very good at following the rulebook, so good that from outside the room it appears that there is a native Chinese speaker on the inside. This, Searle suggests, gives us a system that is analogous to a computer instantiating a program, a system that passes through the same functional states as a native Chinese speaker does when understanding Chinese. But, says Searle, no matter how good he gets at manipulating the squiggles, he does not understand Chinese. His outputs don't mean anything to him; they lack intentionality. Thus, functionalism fails to account for the intentionality of mental states and hence fails to be an adequate theory of mind.

Functionalists have various responses to this argument. One prominent response charges that Searle is looking for understanding in the wrong place. He is just a cog in the machine while it's the overall system of which he is a part that achieves understanding. (See, e.g., Boden 1988) But even if

[17] Ruth Millikan's work has been especially important in the development of the teleological notion of function (see, e.g., Millikan 1993).

[18] For further discussion of intentionality, see Chapter 8.

functionalism is able to account for intentional states like beliefs and desires – and many philosophers think that despite Searle's objections the theory is especially well suited in this regard – it has faced intense criticism regarding its ability to handle qualitative states.

Consider the experience of seeing a ripe banana, or smelling a skunk's spray, or feeling a dull ache in your lower back. Each of these experiences has *phenomenal* or *qualitative* character – to use a phrase associated with the work of Thomas Nagel, there is *something it is like* to have such experiences. The experience of seeing a ripe banana has a different qualitative character from seeing an unripe banana, and the experience of feeling a dull ache in your lower back has a different qualitative character from the experience of feeling a sharp twinge in your lower back.

Two different arguments have been offered to show that functionalism cannot adequately account for the qualitative character of our mental states. The first such argument – typically referred to as the *absent qualia argument* – owes primarily to the work of Ned Block (1978). Block proposed a thought experiment involving a homunculi-headed robot, i.e., a robot whose body is powered by a system consisting of a billion homunculi.[19] Supposing we're able to map out the functions of the human brain in a machine table, we could assign each homunculus a simple task corresponding to one square of that table, e.g., pushing a certain output button upon receiving a certain input. In this way, the billion homunculi together would constitute a system that is functionally equivalent to the human brain. According to Block, however, it seems implausible that such a system would really feel pain or have the qualitative experience associated with seeing a ripe banana. To demonstrate this implausibility, Block proposes that we recruit one billion humans and have each of them substitute for one of the homunculi. When thinking about a robot powered in this way, most people have the strong intuition that it would lack qualia. But since having qualitative character is essential to the mental state of pain, and to the mental state of seeing a ripe banana, functionalism does not provide an adequate account of these states.

The second qualia-based argument directed at functionalism is what's typically referred to as the *inverted qualia argument*. Underlying the argument is the intuition, first articulated in the 17th century by John Locke, that inversion of the visible spectrum might be behaviorally undetectable, i.e., that two people might have quite different – even inverted – qualitative experiences without this difference showing up in their behavior. Starting in the 1970s, several philosophers began employing the possibility of spectrum inversion in arguments against functionalism (see, e.g., Block and Fodor 1972, Shoemaker 1975). The argument goes roughly as follows. Consider two individuals, Ruby and Kelly, who are functionally identical to one another with respect to their color experiences. Both will refer to red tomatoes as ripe and to green tomatoes as unripe; both stop at red lights and go at green lights; both note that a stop sign has the same color as a Coke can, and that grass has the same color as Kermit the frog. But it seems possible that their qualitative experiences are very different from one another. The experience that Ruby has when looking at a ripe tomato might be different from the

[19] Block supposed that a billion homunculi would be sufficient to realize the functional organization of the human brain since, at the time that he was writing, that corresponded to the best estimate for the number of neurons in the brain. It is now believed that there are upwards of 85 billion neurons in the brain. Block can of course accommodate this development by increasing the number of homunculi needed to power the robot system.

experience that Kelly has when looking at a ripe tomato. In particular, Kelly's experience when looking at the ripe tomato might be the experience that Ruby has when looking at Kermit the frog, while the experience that Kelly has when looking at Kermit the frog might be the experience that Ruby has when looking at a ripe tomato. In this case, though Ruby and Kelly have experiences that are inverted from one another, there will be no functional difference. But since the qualitative aspect of a mental state seems central to its being the mental state that it is, functionalism seems inadequate.

Granted, many philosophers have questioned the coherence of spectrum inversion (see, e.g., Dennett . But accounting for qualia has continued to prove problematic for functionalism and in fact, for physicalism more generally. In the last quarter of the 20th century, debates about qualia emerged to play a key role in the mind-body problem. We turn to these debates in the next section.

4. The Age of Qualia

With respect to the mind-body problem, the end of the 20th century can in many ways be thought of as the age of qualia. Although the majority of philosophers in the 21st century still consider themselves to be physicalists of one sort or another (see Bourget and Chalmers 2014), since the 1970s there has been significant philosophical attention devoted to qualia and, in particular, to the apparent difficulty in accounting for qualia within a physicalist treatment of mind.[20] If 20th century discussion of the mind-problem began with the Age of Behaviorism, and subsequently passed through Age of the Identity Theory and the Age of Functionalism, it would not be much of an overstatement to characterize the end of the century (and indeed, the beginning of the 21st century) as the Age of Qualia.

In addition to the qualia-based arguments against functionalism that we considered in the previous section, two related arguments that emerged in the 1970s and 80s brought qualia to the forefront of discussion. The first of these arguments has become known as *the bat argument*; the second has become known as *the knowledge argument*. In the 1990s, a third qualia-based argument – *the zombie argument* – entered the fray.[21] All three of these arguments aim to show that physicalism cannot adequately account for qualia and thus cannot be an adequate theory of mind. At the same time, they have also had the effect of rejuvenating the dualist position that had been dormant for most of the century. In what follows we consider each of these arguments in turn.

4.1. The Bat Argument

Thomas Nagel introduced the bat argument in his seminal article "What Is It Like to Be a Bat?" (1974). Given that bats are mammals, they are surely conscious – there is surely something that there's like to be a bat. But bats navigate the world very differently from the way that we humans do. While we use our senses of sight, sound, and touch to make our way about the world, bats do so by way of echolocation. Thus, their conscious experience is very different from ours – so different, in fact, that we cannot even imagine what it's like for a bat when it is using its sonar. What it's like to be a bat is thus

[20] This debate has been accompanied by a corresponding surge of interest in the notion of consciousness more generally. For further discussion of consciousness, see Chapter 3.

[21] An additional influential critique of physicalism related to qualia-based considerations was developed in Kripke 1980.

fundamentally a subjective phenomenon, understood only from a single point a view (namely, the bat's). Since physicalism takes the objective point of view, it cannot capture what it is like to be a bat. Moreover, this failure of physicalism is not a minor one, since the fact that experience is subjective is an essential fact about experience, i.e., the subjectivity of what it is like to be a bat is an essential fact about it. So, concluded Nagel, physicalism cannot capture all the essential facts about experiences.

Even though our own conscious experience is very different from that of the bat, one might question whether Nagel was right to conclude that we can't even imagine it. In an attempt to forestall this kind of objection, Nagel noted that it wouldn't be enough for us to imagine that we have webbing on our arms, that we spend the day in caves hanging upside down by our feet, that we eat insects, and so on. It won't even help to imagine that one has extremely poor vision and that one uses high-frequency signals to perceive the world. As Nagel argued:

> In so far as I can imagine this (which is not very far), it tells me only what it would be like for *me* to behave as a bat behaves. But that is not the question. I want to know what it is like for a *bat* to be a bat. Yet if I try to imagine this, I am restricted to the resources of my own mind, and these resources are inadequate to the task. I cannot perform it either by imagining additions to my present experience, or by imagining segments gradually subtracted from it, or by imagining some combination of additions, subtractions, and modifications. (Nagel 1974, 439)

Since the time Nagel wrote his article, developments in virtual reality have made possible human experience of something like echolocation. But presumably Nagel would extend the reasoning from the above quotation to deny that even this experience would be enough to enable us to imagine what it is like to be a bat: All that we could learn from such an experience would be what it is like for a human to have some bat-like qualities.

4.2. The Knowledge Argument

The knowledge argument was introduced by Frank Jackson in the 1980s in a pair of articles: "Epiphenomenal Qualia" (1982) and "What Mary Didn't Know" (1986). The argument centers around a thought experiment involving Mary, a brilliant color scientist.[22] We are asked to imagine that Mary has lived her entire life enclosed in a black and white room and that she has never been exposed to color. She wears black and white gloves, she never presses on her eyeball to have a phosphene experience, and so on. While in the room, however, she has been given black and white textbooks, a black and white television, a computer with a black and white monitor, and other black and white research tools. Moreover, Mary lives at a future time at which researchers have developed a completed color science. While in her black and white room, Mary masters this color science. Through careful study, that is, she learns the entire physical story of color. She knows exactly how the human eye and the human brain process color, she knows exactly how humans categorize objects by color, and she knows about the similarity relations among colors. Now suppose that one day Mary is released from her black and white

[22] Jackson 1982 also included a second thought experiment involving a man named Fred who's able to discriminate more colors than normal human beings. Subsequent discussion, however, has tended to focus almost exclusively on the Mary case.

environment. Immediately upon her release, Mary sees a ripe tomato. According to Jackson, it seems overwhelmingly plausible that this experience provides Mary with an "aha" moment: Once she sees color for the first time, she learns something new. In particular, she learns what seeing red is like. But since she already knew all of the physical facts about color, since she already knows the entire physical story, that story must not be the whole story. For this reason, physicalism cannot provide an adequate account of our mental states.

Physicalist responses to the argument tend to divide into two broad categories. First, some physicalists have denied Jackson's intuition about Mary, what we might call the *'aha' intuition*. According to Daniel Dennett, who has persistently pushed this line in response to Jackson, the reason that we mistakenly have the 'aha' intuition is that we've merely managed to imagine that Mary has lots and lots of physical information, not that she has *all* the physical information. If we were to imagine the situation correctly – if we were really to imagine that Mary has *all* the physical information – we would see that there is nothing left for Mary to learn. (Dennett 1991, 398)

This strategy has not been widely pursued, presumably because even most physicalists find the 'aha' intuition very hard to deny. Most physicalists instead pursue a second strategy, one that concedes that Mary learns something new upon leaving the room. Such philosophers deny that this concession threatens physicalism, because they deny that what Mary learns consists of a new fact. This second kind of response to the Mary case itself divides into two broad classes. The first group of philosophers deny that Mary's newfound knowledge is factual. Rather, it is a different kind of knowledge – perhaps know-how (Lewis 1990; Nemirow 1990), or perhaps acquaintance knowledge (Conee 1985). A second group of philosophers accept that Mary's knowledge is indeed factual, but they deny that it's knowledge of a new fact. Rather she comes to recognize an old fact in a new way, under a new guise or via new concepts (Loar 1990). Despite such responses, however, the knowledge argument has continued to have considerable traction in philosophy of mind.[23]

4.3. The Zombie Argument

The zombie argument came to prominence in the mid-1990s through the work of David Chalmers.[24] The philosopher's zombie is importantly different from the brain-eating creatures that populate Hollywood horror movies. As described by Chalmers in *The Conscious Mind* (1996), zombies are understood to be creatures who are physically identical to human beings but who completely lack phenomenal consciousness, i.e., they are completely lacking in qualia. Your zombie twin, for example, is molecule-for-molecule identical to you and likewise identical to you functionally: She processes information just as you do, reports on her mental states just as you do, focuses her attention on the world just as you do, and so on. But, as Chalmers said, "none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel. There is nothing it is like to be a zombie." (Chalmers 1996, 95)

[23] Jackson himself, however, eventually recanted; he no longer believes that the knowledge argument disproves physicalism. See Jackson 2003.

[24] A similar argument was previously introduced in Kirk (1974).

Chalmers then argued from the conceivability of zombies to the falsity of physicalism. If we can conceive of a zombie world – a world that is physically identical to ours yet in which there is a complete absence of phenomenal consciousness – then such a world is metaphysically possible. But if a zombie world is metaphysically possible, then facts about consciousness are facts over and above the physical facts. Since the truth of physicalism requires that there be no facts about consciousness that are over and above the physical facts, physicalism must be false.

The zombie argument is often referred to as a *conceivability argument*. It moves from facts about what's conceivable to facts about what's possible. In this regard it resembles Descartes' famous argument for substance dualism, presented in his *Meditations on First Philosophy* (1642). Descartes rested his argument on the claim that he could conceive of the mind existing without the body; from this he concluded that it is possible for the mind to exist without the body and hence that the mind and the body are two separate substances. Chalmers' conceivability argument does not aim to establish substance dualism but rather the falsity of physicalism. His own positive view, developed subsequently in *The Conscious Mind*, is a naturalistic version of property dualism.

Conceivability arguments typically face two different kinds of objections. First, it might be questioned whether the proposed scenario is really conceivable. Second, it might be questioned whether conceivability is really a good guide to metaphysical possibility, i.e., it might be questioned whether the conceivability of a given scenario really shows that such a scenario is logically possible. (See Gendler and Hawthorne 2002) In addition to criticisms of these sorts, the zombie argument also faces a third kind of criticism. Many physicalists question whether the metaphysical possibility of zombies counts against physicalism, and this in turn leads to questions about how exactly physicalism should be construed. Twenty years after its articulation, the zombie argument continues to be heavily debated.

V. Conclusion

From the vantage point of these early years of the 21st century, it is still too soon to assess whether we've reached the end of the age of qualia, and if so, what new age will be ushered in to replace it. It seems clear, however, that phenomenal consciousness continues to pose a threat to physicalist theories of mind. This threat has led some to argue that the mind-body problem is in principle insoluble (see, e.g., McGinn 1989). It also seems to account for the recent resurgence of interest in theories of mentality that in various ways aspire to transcend the traditional dualism-physicalism divide. In particular, many philosophers of mind have begun to explore the coherence of positions like panpsychism and Russellian monism, both of which try to find a place for consciousness at the fundamental level of reality.[25]

But despite the threat of phenomenal consciousness, traditional versions of physicalism continue to enjoy considerable support. In the view of many physicalists, the tremendous

[25] See, e.g., the collection of papers in Alter and Nagasawa 2015.

neuroscientific progress of the 20th century suggests that it is just a matter of time before we are able to understand mentality entirely in terms of neural mechanisms. Perhaps this will require us to abandon some of our common mental state vocabulary; it might be that our folk psychological concepts like "belief" and "desire" do not map very well unto neuroscientific states. Such is the prediction made by eliminative materialists such as Paul Churchland (see, e.g., his 1981). It might be that we will only be able to understand the truth of physicalism by way of a conceptual revolution of sorts.[26] Alternatively, perhaps, future developments in neuroscience, in conjunction with philosophical theorizing, might enable us better to grasp how a physical reduction of mentality is possible.[27]

Ultimately, however, it remains the case at the start of the 21st century that the nature of mentality is still very much in dispute. Granted, there are some points of widespread agreement. Substance dualism, which had dominated philosophy prior to the 20th century since the time of Descartes, is no longer considered viable. Likewise, behaviorism has been dismissed as a failed experiment. But despite these important points of agreement and the corresponding philosophical progress involved, it is clear that the 20th century did not provide a widely accepted solution to the mind-body problem. It remains to be seen whether such a solution will be found in the century ahead.[28]

References

Torin Alter and Yujin Nagasawa. 2015. *Consciousness in the Physical World: Perspectives on Russellian Monism*. Oxford: Oxford University Press.

Armstrong, David M. 1999. *The Mind-Body Problem: An Opinionated Introduction.* Boulder, Col.: Westview Press.

Armstrong, David M. 1968. *A Materialist Theory of Mind*. New York: Humanities Press.

Boden, Margaret. 1988. *Computer Models of the Mind*. Cambridge: Cambridge University Press.

Boring, Edwin G. 1933. *The Physical Dimensions of Consciousness.* New York: The Century Co.

Bourget, David and Chalmers, David J. 2014. "What Do Philosophers Believe?" *Philosophical Studies* 170: 465-500.

Byrne, Alex. 1994. "Behaviorism." In Samuel Guttenplan, ed., *A Companion to the Philosophy of Mind.* Oxford: Blackwell.

Carnap, Rudolf. 1932. "Psychology in Physical Language." *Erkenntnis* 3: 107-42.

[26] This point has recently been argued by Nagel: "Our inability to come up with an intelligible conception of the relation between mind and body is a sign of the inadequacy of our present concepts." (1998).

[27] For a more detailed discussion of what lies ahead for philosophical theorizing about mentality, see Chapter 12.

[28] Thanks to Frank Menetrez and Julie Yoo for comments on a previous draft.

Chalmers, David J. 1996. *The Conscious Mind*: *In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Churchland, Paul. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78: 67–90.

Conee, Earl. 1985. "Physicalism and Phenomenal Properties." *Philosophical Quarterly* 35: 296-302.

Crawford, Sean. 2013. "The Myth of Logical Behaviourism and the Origins of the Identity Theory." In Michael Beaney, ed., *The Oxford Handbook of the History of Analytic Philosophy*. Oxford: Oxford University Press.

Davidson, Donald. 1970. "Mental Events." In Lawrence Foster and J. W. Swanson, eds., *Experience and Theory.* London: Duckworth.

Dennett, Daniel. 1991. *Consciousness Explained. Boston:* Little, Brown and Company.

Dennett, Daniel. 1988. "Quining Qualia." In A. Marcel and E. Bisiach, eds., *Consciousness in Modern Science*. Oxford: Oxford University Press.

Dennett, Daniel. 1987. *The Intentional Stance.* Cambridge, Mass: The MIT Press.

Feigl, Herbert. 1958. "The 'Mental' and the 'Physical'." *Minnesota Studies in the Philosophy of Science* 2: 370-497.

Feyerabend, Paul. 1963. "Mental Events and the Brain." *Journal of Philosophy* 40:295–6.

Gendler, Tamar and Hawthorne, John. 2002. "Introduction: Conceivability and Possibility." In Tamar Gendler and John Hawthorne, eds., *Conceivability and Possibility*. New York: Oxford University Press.

Graham, George. 2003. "Behaviorism." In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2015 edition). Available at <http://plato.stanford.edu/archives/spr2015/entries/behaviorism/>

Fodor, Jerry. 1981. "The Mind-Body Problem." *Scientific American* 244: 114-25.

Fodor, Jerry. 1968. *Psychological Explanation*. New York: Random House.

Hempel, Carl G. 1980. "The Logical Analysis of Psychology." In Ned Block, ed., *Readings in the Philosophy of Psychology*, Volume 1. Cambridge, Mass.: Harvard University Press, 1-14.

Hill, Christopher S. 1991. *Sensations: A Defense of Type Materialism*. Cambridge: Cambridge University Press.

Jackson, Frank. 2003. "Mind and Illusion." In Anthony O'Hear, ed., *Minds and Persons*. Cambridge: Cambridge University Press. 421-442.

Jackson, Frank. 1986. "What Mary Didn't Know." *Journal of Philosophy* 83: 291-5.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127-136.

James, William. 1890/1981. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Kim, Jaegwon. 2011. *Philosophy of Mind* (third edition). Boulder, Col.: Westview Press.

Kim, Jaegwon. 1992. "Multiple Realization and the Metaphysics of Reduction." *Philosophy and Phenomenological Research* 52: 1–26.

Kirk, Robert. 1974. "Zombies Vs Materialists." *Proceedings of the Aristotelian Society* 48: 135-52.

Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Levin, Janet. 2013. "Functionalism." In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Fall 2013 edition). Available at <http://plato.stanford.edu/archives/fall2013/entries/functionalism/>

Lewis, David. 1990. "What Experience Teaches." In William G. Lycan, ed., *Mind and Cognition*. Oxford: Blackwell, 29-57.

Lewis, David. 1980. "Mad Pain and Martian Pain." In Ned Block, ed., *Readings in the Philosophy of Psychology*, Volume 1. Cambridge, Mass.: Harvard University Press, 216-222.

Lewis, David. 1966. "An Argument for the Identity Theory." *Journal of Philosophy* 63: 17–25.

Loar, Brian. 1990. "Phenomenal states." *Philosophical Perspectives* 4: 81-108.

McGinn, Colin. 1989. "Can We Solve the Mind-Body Problem?" *Mind* 98: 349-66

Millikan, Ruth. 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Nagel, Thomas. 1998. "Conceiving the Impossible and the Mind-Body Problem." *Philosophy* 73: 337-52.

Nagel, Thomas. 1974. "What Is It Like To Be a Bat?" *Philosophical Review* 83: 435-50.

Nemirow, Laurence. 1990. “Physicalism and the Cognitive Role of Acquaintance.” In William G. Lycan, ed., *Mind and Cognition*. Oxford: Blackwell.

Papineau, David. 2002. *Thinking About Consciousness*. Oxford: Oxford University Press.

Place, Ullin T. n.d. “Identity Theories.” *A Field Guide to the Philosophy of Mind.* Available at *http://host.uniroma3.it/progetti/kant/field/mbit.htm*

Place, Ullin T. 2002. “A Pilgrim’s Progress? From Mystical Experience to Biological Consciousness.” *Journal of Consciousness Studies* 9: 34-52.

Place, Ullin T. 1956. “Is Consciousness a Brain Process.” *British Journal of Psychology* 47: 44-50.

Puccetti, Roland. 1977. “The Great C-Fiber Myth: A Critical Note.” *Philosophy of Science* 44: 303-305.

Putnam, Hilary. 1975. *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press.

Putnam, Hilary. 1967. “The Nature of Mental States.” Originally published as “Psychological Predicates” in Capitain and Merrill, eds., *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press. Reprinted in Putnam 1975, 429-440. Page references to reprinted edition.

Putnam, Hilary. 1963. “Brains and Behavior.” In R. Butler, ed., *Analytical Philosophy Second Series* (Oxford: Basil Blackwell). Reprinted in Putnam 1975, 325-341. Page references to reprinted edition.

Quine, Willard van Orman. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Rorty, Richard. 1965. “Mind-Body Identity, Privacy, and Categories.” *Review of Metaphysics* 19: 24-54.

Schlick, Moritz. 1925/1974. *General Theory of Knowledge.* Translated by Alfred E. Blumberg. New York: Springer-Verlag.

Sellars, Wilfred. 1956. “Empiricism and Philosophy of Mind.” In H. Feigl and M. Scriven, eds., *The Foundations of Science and the Concepts of of Psychology and Psychoanalysis: Minnesota Studies in the Philosophy of Science*, Vol. 1. Minneapolis: University of Minnesota Press.

Skinner, B.F. 1974. *About Behaviorism*. New York: Alfred A. Knopf, Inc.

Smart, J.J.C. 1959. “Sensations and Brain Processes.” *Philosophical Review* 68: 141-56.

Titchener, E.B. 1914. “On ‘Psychology as the Behaviorist Views It.’” *Proceedings of the American Philosophical Society* 53: 1-17.

Watson, John B. 1913. "Psychology as the Behaviorist Views It." *Psychological Review* 20: 158-77.