APPENDIX I

Some Backdrop

Dualism, Behaviorism, Functionalism, and Beyond

The present text begins quite close to where most philosophical treatments end: with recent attempts to understand mindfulness using the tools of neuroscience, cognitive psychology, and artificial intelligence. In these brief notes[1] I offer some rough-and-ready background, in the form of a few cameos of a few historically important positions.

1. Dualism

When we introspect, or reflect on our own thoughts, feelings, and beliefs, we do not find anything much like physical objects. Beliefs may be important or trivial, feelings strong or weak, but not literally big, or colored, or heavy, and so on. On the evidence of introspection alone, then, we might be inclined to conclude that the mind is something quite separate from, and deeply distinct from, the physical world. This perfectly natural viewpoint is known as *dualism.*

Considered as a philosophical theory of the nature of mind, Dualism is somewhat uninformative. It tells us what the mind is *not*; it is not a normal physical item like a body, brain, table, or chair. But it is embarrassingly silent about what it might actually *be.* But still, knowing that the moon is not made of green cheese is quite handy, even if you do not know what it is actually made of instead. So let us begin by giving the dualists' claim—that the mind is not a physical item—the benefit of the doubt. The question then arises: What is the relationship between this nonphysical item and the physical body that accompanies it around the world?

[1]These notes are based on some of my longstanding classroom teaching materials, and in one or two places I wonder whether something might have been unwittingly borrowed from some other source. My best efforts at checking this reveal no such unacknowledged borrowings. But should something have slipped the net of appropriate citation, I hereby apologize: and do let me know!

When dualism was in its heyday, around the time of the seventeenth century, there were three major contenders as an account of this relation:

1. Parallelism
2. Epiphenomenalism
3. Interactionism

1. According to the parallelist, the mind and the body are distinct and causally isolated. Neither is capable of affecting the other. How, then, are we to account for the *appearance* of causal linkage; the impression we have of wishes causing action and blows to the head causing hallucinatory experiences? Synchronization was to be the key. God, or some other force or agency, had arranged matters so that the two causal orders—the mental and the physical—would run along in harmony, like two ideally accurate clocks set to the same initial time and left to run for eternity; neither sustaining or consulting the other, but the two in perfect accord nonetheless.

The trouble with parallelism is who set the clocks? And why, if it was God, did God resort to such a clumsy piece of trickery?

2. *Epiphenomenalism* is like parallelism in asserting the causal isolation of the physical from the mental. But it relaxes the requirement in the other direction. The epiphenomenalist allows that the physical can cause the mental, but denies that the mental can affect the physical. The mind, on this account, is somewhat (though only somewhat) like the exhaust fumes from a car. The fumes accompany and are caused by the activity of the engine. But they do not (typically) power the car. Just so, the epiphenomenalist holds that beliefs and thought and other mental experiences accompany and are caused by brain activity. But they do not actually cause the body to act. They are just the icing on the cognitive cake. This is a counterintuitive prospect indeed; it certainly *feels* as if it is my desire for a Pete's Wicked Ale that prompts the trek to the local hostelry. Insofar as the whole impetus for accounts that reserve a special place for mental phenomena comes from a desire to respect the introspective evidence, this seems an odd conclusion to have to accept.

3. *Interactionism* is the most immediately appealing of the dualist positions. It treats the mental and the physical as distinct but causally integrated items, thus avoiding some of the metaphysical excesses and introspective implausibility of parallelism and epiphenomenalism. The most famous form of interactionism is Cartesian dualism. On Descartes' famous model, the mind is a totally nonphysical substance that acts on the body by influencing the pineal gland at the base of the neck. The body, by the same route, influences the mind.

The problem most commonly urged against Cartesian dualism is: How do two such distinct items—the body and the mind—manage to be parts of a single causal network? We understand, we think, how the physical can affect the physical; but how can the nonphysical do so?

The argument has some force. Cartesian dualism would certainly gain in plausibility if we had some such account. Still, we allow that many things that are not at all *like* physical objects may still act on them. Witness (to take a classic case) the iron filings acted on by a magnetic field. So it is not obviously the case that Cartesian interactionism is *conceptually* impossible.

So why give up dualism?

Dualist doctrines of the kind outlined above have been largely abandoned by science and philosophy. The mind is now taken to be grounded in the physical body in such a way that the problem of interaction need not arise. Many factors have contributed to dualism's downfall. Probably the most important of these are the following.

1. The obvious *dependence* of the mental on the physical. Drugs (such as Prozac, or ecstasy), which affect the physical constitution of the brain in moderately well understood ways, systematically affect our moods and emotions. Brain damage—for example, an iron spike through the prefrontal cortex—is likewise disruptive. The evolution of intelligent creatures is correlated with changes in brain structure. All this suggests (as presented in Churchland, 1984) that we must *at least* look for a systematic correlation of brain activity and mental activity. Why, then, assume that there are two *items* here, in need of correlation, instead of one item exhibiting a variety of properties? Materialism—the thesis that we are dealing with just one kind of *item* or substance, viz. physical matter—seems to win out on grounds of simplicity.

2. The *positive* arguments in favor of dualism are unconvincing. These are (a) the "how could . . . ?" argument, and (b) the argument from introspection.

   a. The "how could . . . ?" argument relies on finding properties of human beings and asking "Now how could any mere *physical* system do *that*?" Descartes suggested that *reasoning* and *calculation* were beyond any mere physical system. But today, with our intuitions molded by shops full of Palm Pilots, G4s and even modest pocket calculators, we are unlikely to choose calculation to fill in the blank. Now people are more likely to choose some ability like "falling in love," "appreciating a symphony," or "being creative." But work in neuroscience and artificial intelligence is steadfastly eroding our faith that there are some things that no mere physical system could ever do. As such, the fact that *we* do *X*, *Y*, or *Z* no longer cuts much ice as an argument to the effect that we *cannot possibly* be "mere" physical systems.

   b. The argument from introspection is a harder nut to crack. The idea is that we *just know* that a belief is not a state of brain or body. We can tell just by looking "inside ourselves" and seeing what a feeling is *like*. The trouble here is that introspection is a weak kind of evidence. Granted, we know that our feelings do not *strike us* as being brain states. But so what? I may have a feeling in my stomach that does not strike me as being a mild case of salmonella. But it might still *be* a mild case of salmonella for all that. This oversimplifies the issue, but the general point is clear. Unless someone can show that what introspection reveals cannot be the *very same thing* as a bodily state, albeit under a different description, we need not accept introspection as decisive evidence in favor of dualism.

Dualism, then, lacks explanatory force and independent positive evidence in its favor. How else might we conceive the mind?

2. Behaviorism

Probably the first major philosophical reaction against Dualism came not as a result of the explanatory inadequacies just described, but instead grew out of a movement within philosophy that is sometimes referred to as the *linguistic turn*. The leading idea was that philosophical puzzles were at root puzzles about *language*. Gilbert Ryle, in *The Concept of Mind*, published in 1949, accuses dualism and the whole body–mind debate of a failure to understand the role of mental talk in our language. Philosophy of mind, according to Ryle, was captivated by *Descartes' myth*. And Descartes' myth was, in effect, the idea of mind as an inner sanctum known only by introspection. The myth inclined philosophers to seek some account of the relation of this inner sanctum to the public world of people, objects, and actions. But the task was thought to be misconceived. Philosophers, Ryle claimed, were failing to see the significance of mental talk, in much the same way as someone fails to see the significance of talk about a university who, on being shown the library and colleges and playing fields and accommodation, goes on to complain, "Yes. I see all that. But where is *the university*?" The answer is that the university is not something extra, beyond all the colleges, accommodation, and so on. It is just the organization of those very items. Just so, Ryle argued, the mind is not something beyond all its public behavioral manifestations—mindtalk is just a way of talking about the organization of the behavior itself. When we say that Mary loves teaching, we do not mean that inside Mary there is a ghostly loving that accompanies her professional acts. Rather we mean only that Mary's actual and potential behavior will follow a certain pattern. That pattern might be expressed as a very long conjunction of claims about what Mary would do in certain situations, e.g.,

if she is offered a new textbook she will take it;

if someone asks her if she likes teaching, she will say yes;

if she sees a good teacher in action, she will try to emulate them

and so on.

The idea, in short, is that mental talk picks out *behavioral dispositions*. It isolates what so and so is likely to do in such and such circumstances. It does not pick out a state of an inner mental sanctum. The classic analogy is with chemical dis-

positions such as solubility. To say that *X* is soluble is not to say that *X* contains some hidden spirit of solubility. It is just to say that if you put *X* in water, *X* would dissolve. Mental talk picks on more complex dispositions [what Paul Churchland (1984) calls "multi-tracked dispositions"]; but dispositions is still *all* they are.

Three worries afflict behaviorism in the form I have presented it.

1. The dispositional analysis looks either *infinite* or *circular*. It will be infinite if we have to list what a given belief will dispose an agent to do in *every possible situation* they could be in. And it will be circular if our list of dispositions makes irreducible reference to other mental states, e.g., Mary will try to teach well as long as she is happy and does not believe teaching is ruining her life.
2. The dispositional account seems to want to rule out the inner sanctum completely. But isn't there some truth in the idea? Don't we have inner feelings, pains, images, and the like?
3. It is *explanatorily shallow*. It tells us, at best, something about how we use mental concepts. But this need not be the end of the story of mind. Even if "soluble" just *means* "would dissolve in water," we can ask after the *grounds* of the disposition to dissolve. We can ask *how* it is possible for something to dissolve in water. So too we may ask how it is possible for someone to love teaching. And the explanation should appeal to a range of facts beyond the surface behavior of the teacher. Indeed, taken at face value, behaviorism seems to commit a kind of "method actors fallacy" (see Putnam, 1980), attributing genuine neural states (of, say, pain) to anyone exhibiting appropriate behavior, and denying pain to anyone able to suppress all the behavioral and verbal expressions of pain.

3. Identity Theory

In the mid to late 1950s philosophers began to realize—or rediscover—that there was more to philosophical life than the analysis of the concepts of ordinary language. Philosophy could, for example, contribute to the study of mind and mental mechanisms by examining the conceptual coherence of scientific theory *schemas*. By this I mean, not examining a particular, well worked out scientific theory in say, neurophysiology, but by considering the intelligibility and implications of general types of scientific account of the mind. One such account—the topic of this section—was the so-called Mind–Brain identity theory. The schema here in brief was mental states *are* brain processes.

This schema was advocated, discussed, and refined by philosophers such as U. T. Place, J. J. C. Smart, and D. Armstrong [see the collection edited by V. C. Chappell (1962) for some of the classic contributions]. The philosophical task, then, is not to decide *whether or not* mental states are brain processes. That is a job for ordinary science. Rather, it is to consider whether this general theory schema is one that is even *possibly* true. Does it even make sense to suppose that thoughts, beliefs, and sensations could be identical with brain processes?

Reasons to doubt that it does include

1. Leibniz' law problems
2. species-chauvinism objections.

Leibniz' law states that if two descriptions pick out the same object, then whatever is true of the object under one description must be true of it under the other. Thus, if Spiderman really is Peter Parker, then whatever is true of Spiderman must be true of Peter Parker, and vice versa. If Aunt May is Peter Parker's ailing relative, then she must be Spiderman's ailing relative also. If Spiderman clings to ceilings, then Peter Parker must cling to ceilings also. Formally,

$$(X)\ (Y)\ [(X = Y) \rightarrow (F)\ (FX \leftrightarrow FY)]$$

Whatever their opinion about Spiderman, many philosophers were unable to see how the mind–brain identity thesis could live up to the Leibniz' law requirement. For consider

- [Spatial location] A brain state may be located in space, say 10 cm behind my eyeball. But it surely won't be true of any mental state—say, my belief that Mark McGuire plays for the Cardinals—that it is 10 cm behind my eyeball.
- [Truth value] A belief may be true or false. But how can a brain state be true or false?
- [Sensational content] A pain may be sharp or tingly. But could a brain state be sharp or tingly?
- [Authority] I seem to have some authority over my mental states. If I sincerely believe I am in agony, it looks as if I must be right. But I do not seem to have any authority over my brain states; a neurophysiologist could surely correct me with regard to those.

One way of responding to these objections is simply to grasp the nettle we are offered and say, "It may not *seem* as if brain states can be true or false, or mental states located in space, but they *are*." It does not seem as if a flash of lightning is an electrical discharge, but it is. And if you have some authority when it comes to spotting flashes of lightning, then you have it when it comes to spotting some kinds of electrical discharge whether you know it or not. The idea behind this kind of response is that Leibniz' law is unreliable in contexts that involve people's *beliefs* about properties of objects, rather than just the *actual* properties of the objects. To once again adapt a strategy used by Paul Churchland (1984), we can display the problem by constructing the following clearly fallacious argument:

1. Mary Jane Watson believes that Spiderman is a hero.
2. Mary Jane Watson does not believe that Peter Parker is a hero.

so,

3. By Leibniz' law—Peter Parker is not identical with Spiderman.

Identity theory thus survives the Leibniz' law crisis. Historically, it succumbed (although sophisticated revivals are increasingly popular today) to a very different kind of objection [first raised by Hilary Putnam (1960) in a series of papers beginning with "Minds and machines"]. The objection is one of *species-chauvinism.* On a *strong* reading of the identity theorists' claims it looks as if *types* of mental state (e.g., being happy, angry, seeing blue, believing that Reagan is dangerous) are now being identified with types of brain state (e.g., the firing of a certain group of neurons, or C-fibers, or whatever). But this claim, on closer examination, looks distinctly implausible. For consider one example.

Suppose we type-identify, say, being in pain with having C-fibers 1–9 firing. Then it follows that *no being without C-fibers can be in pain.* But this seems a very rash, even imperialistic, claim. Might we not encounter extraterrestrial beings who look clearly capable of feeling pain (they wince and groan and so on) yet *lack* C-fibers? Maybe many animals to which we happily ascribe psychological properties such as feeling hungry or angry lack C-fibers, too. Maybe we will soon build intelligent computer systems that have neuromorphic VSLI chips instead of neurons. Must we simply *rule out* the possibility that all these different kinds of physical systems may share some of our psychological states? Surely not. Suppose we discovered that various human beings had different kinds of brain structure, such that when Fred felt pain C-fibers 1–9 fired, but when Andy felt pain D-fibers 1–7 fired. Psychological ascriptions seem almost *designed* to class together different brain states in virtue of their common role in determining types of behavior. Strong type–type identity theory does no justice to this capacity for generalization, and can seem species-chauvinistic as a result.

One way out is for the identity theorist to claim that each individual occurrence of a mental state is identical with some brain state. This is the "token" version of identity theory, so named because it associates *tokens* (individual occurrences) of mental events with brain events, without making claims about the identity of types of mental event with types of brain event. One trouble with this as it stands is that it is explanatorily weak; it leaves us unenlightened as to *why* any particular physical state should be identical with the particular mental event with which it is. One way to remedy this is to build on the idea that psychological ascriptions are in part designed to group together physically disparate brain states in virtue of their common *role* in determining behavior, but to build on it in such a way as to avoid the behaviorist's mistake of *identifying* the psychological state with the outward behavior. This is exactly what Putnam did and the result was another philosophical schema for a scientific theory of mind, viz. *functionalism.*

4. Machine Functionalism

The first wave identity theorist faced a hopeless task, akin, as Daniel Dennett has pointed out, to finding a purely physical account of what all clocks, say, have in common. We would find no useful description, in the language of physics, of the commonality in virtue of which a sundial, a clockwork alarm, and a quartz digital alarm are all said to be *clocks.* What unites these disparate physical objects is the purpose, function, or use that we assign to them. Just so, it seems, there need be no useful physical description that captures what my anger, the dog's anger, the Martian's anger, and the robot's anger all have in common. In some sense it looked to be the functionality of the different physical states that realize our several angers that unites the states *as* angers. Hence, *functionalism* is a schema for a scientific theory of mind.

One way of understanding the functionalist approach is by analogy with computer programs. A program is just a recipe for getting a job done, and can be specified, at a very abstract level, as a set of operations to be performed on an input and yielding a certain output—maybe a number or sentence. Defined at such an abstract level the same program can be written in different high-level languages (BASIC, PASCAL, LISP, JAVA, or whatever) and run on machines with very different kinds of hardware. The abstract idea of a program (its input-inner operations–output profile) is captured in its specification as a *Turing machine* (see Chapter 1), which is, in effect, just a description of a fixed set of operations to be performed on whatever strings of symbols it is given as input. The point is that this abstract notion of a program is not "hardware-chauvinist"; the same program, so defined, may run on lots of different physical machines. The functionalist claim, in effect, is that the mind is to the body/brain as the program is to the physical machine.

The analogy is so satisfying, indeed, that the original functionalists went further and claimed not just

C1 The mind is to the brain as the program is to the machine, but

C2 The mind *is* a program, run (in humans) with the brain as its supporting hardware.

C2 is often called *machine functionalism.* Since much of the present text is concerned with versions of machine functionalism, I shall not pursue this position any further here.

4. Eliminativism

The task so far has been to see what general kind of schema for a scientific theory could make sense of the relation between our talk of the mind and some kind of description (functional, behavioral, or whatever) of the physical world. The question was thus:

> What *kind* of scientific theory could possibly count as a theory of the mind?

Some would regard this as a mistaken goal. For it seems to assume that our commonsense ideas about mental phenomena, which together make up our commonsense idea of *mind,* are (at least largely) correct. It assumes, in effect, that there

really are such things as hopes, desires, fears, beliefs, and so on, and that the job of science is to *explain* them. But, after all, people once thought that there were ghosts and vampires and that apparently empty space was filled by mysterious ether and much else that science has shown to be misguided. Imagine, then, a discipline devoted to investigating what *kind* of scientific theory could possibly account for the existence of the ether. What a waste of time! What science shows is that there is no ether and so the task of accounting for its existence never arises. Could the commonsense notion of mind meet a similar fate? Those who think so call themselves *eliminative materialists* (e.g., Churchland, 1981). The task of philosophy, as they see it, is not to prejudge the issue by simply setting out to discover what scientific schema explains the commonsense view of mind, but also to critically examine scientific accounts to see whether the commonsense view is *sound.* Once again, this is a topic treated in the main text and I shall not pursue it far here. Notice, however, that eliminative materialism need not be an all or nothing doctrine. Dennett (1987), for example, allows that some of our common sense ideas about the mental may find a home in some future scientific theory. He just denies that we should *demand* that any good theory capture all our pretheoretical intuitions.

The most radical versions of eliminative materialism predict that virtually nothing of the commonsense framework will be preserved. Beliefs, desires, hopes, and fears will all be abandoned in some future science of the mind. It is, I suspect, extremely hard to even make sense of this claim *in advance* of the science being developed and offering us alternative concepts to use when we formulate it. From here, it is hard to see how such a future science could *be* a science of the mind at all. But that, of course, may just be predictable conceptual myopia. On the other hand, it does seem as if there is a whole cluster of related concepts involving actions, beliefs, and desires that just *constitute* the idea of mind. We could certainly give some up and revise others. But could we really drop them all? And to what extent does the legitimacy of those concepts depend on their finding a place in some scientific theory anyway? It is a virtue of eliminative materialism that it is radical enough to bring these issues to the fore.

Suggested Readings

Several recent textbooks offer superb introductions to the topics covered in this appendix. I especially recommend J. Kim, *Philosophy of Mind* (Boulder, CO: Westview, 1996) and D. Braddon-Mitchell and F. Jackson's *Philosophy of Mind and Cognition* (Oxford, England: Blackwell, 1996). Other useful treatments include G. Graham, *Philosophy of Mind: An Introduction* (Oxford, England: Blackwell, 1993) and P. Churchland's classic, *Matter and Consciousness* (Cambridge, MA: MIT Press, 1984, and many subsequent and expanded editions). W. Lycan (ed.), *Mind and Cognition: A Reader* (Oxford, England: Blackwell, 1990) offers a fine collection of papers covering functionalism, identity theory, eliminativism, and much else besides.

APPENDIX II

Consciousness and the Meta-Hard Problem

Readers of some early versions of this text suggested that it paid too little attention to the hot topics of consciousness and subjective experience. This was no accident. But it is undeniably the case that a complete and satisfying scientific account of the nature of mindware cannot remain forever silent concerning what is, arguably, the single most puzzling fact about mind! It is with some trepidation, then, that I offer a sketch of the issues (as they appear to me) and a few critical and constructive remarks.

Consciousness has certainly come out of the closet. After a long period during which the word was hardly mentioned in scientific circles, consciousness is now the star of a major growth industry. There are books, meetings, and journals. There are Internet discussion groups and web sites. There is hope, interest, and excitement. But is there a theory—or even a promising sketch for a story? It is, strangely, rather hard to say. It is hard to say because first, the word "consciousness" does not seem to aim at a single, steady target. We need to distinguish various possible targets and assess the state of the art relative to each one. And second, it is unclear (especially with respect to some of the more recondite targets) exactly what would *count* as a theory, sketch, story, or explanation, anyway.

Some possible targets for a theory of consciousness include simple awakeness, self-awareness, availability for verbal report, availability for the control of intentional action, and, of course, the star of the show—raw feels or qualia, the distinct feels and sensations that make life worth living or (sometimes) worth leaving.

Simple awakeness may be roughly defined as the state in which we are quite sensitive to our surroundings, able to process incoming information and respond appropriately. Self-awareness involves the capacity to represent ourselves and to be aware of ourselves as distinct agents. Availability for verbal report involves both a capacity to somehow access our own inner states and to describe what we find using words (or sign language, etc.). Availability for the control of intentional action