FOURTH EDITION

METAPHYSICS

PETER VAN INWAGEN

University of Notre Dame

WESTVIEW PRESS

A MEMBER OF THE PERSEUS BOOKS GROUP

Westview Press was founded in 1975 in Boulder, Colorado, by notable publisher and intellectual Fred Praeger. Westview Press continues to publish scholarly titles and high-quality undergraduate- and graduate-level textbooks in core social science disciplines. With books developed, written, and edited with the needs of serious nonfiction readers, professors, and students in mind, Westview Press honors its long history of publishing books that matter.

Published by Westview Press,
A Member of the Perseus Books Group

Find us on the World Wide Web at www.westviewpress.com.

Westview Press books are available at special discounts for bulk purchases in the United States by corporations, institutions, and other organizations. For more information, please contact the Special Markets Department at the Perseus Books Group, 2300 Chestnut Street, Suite 200, Philadelphia, PA 19103, or call (800) 810-4145, ext. 5000, or e-mail special.markets@perseusbooks.com.

Cover design by Miguel Santana and Wendy Halitzer
Book design by Cynthia Young

Library of Congress Cataloging-in-Publication Data
Van Inwagen, Peter.
Metaphysics / Peter van Inwagen, University of Notre Dame.
— FOURTH EDITION.
pages cm
Includes bibliographical references and index.
ISBN 978-0-8133-4934-3 (pbk.) — ISBN 978-0-8133-4935-0 (e-book)
1. Metaphysics. I. Title.
BD111.V38 2014
110—dc23

2014012814

10 9 8 7 6 5 4 3 2 1

PART THREE

THE INHABITANTS OF THE WORLD

INTRODUCTION TO PART THREE

The final part of this book is about *us*, the inhabitants of the World. That is, it is about us human beings and any other beings there may be that are sufficiently similar to us that it would be reasonable to consider them our fellow inhabitants of the World. (While it may be reasonable to use the word 'inhabitants' in a sense in which apes and beavers and elephants—and perhaps even ants—are "our fellow inhabitants of the World," I will use the word in the sense suggested by the adjective 'inhabited'—as in the question "Is that island inhabited?") The term traditionally used to describe us and beings "sufficiently similar" to us is 'rational'. Human beings, however irrationally they may behave, and angels and Martians (if there are angels or Martians) are rational in the required sense. Apes and beavers and elephants are not rational in the required sense.[1] Non-human terrestrial animals—especially apes—may, however, be very *intelligent*. For this reason, in Part Three, I avoid using the term 'intelligent' to do the work I now assign to the word 'rational'. The use of 'intelligent' and 'intelligence' to refer to mental capacities not possessed by even the brightest apes is quite common, as may be seen from such familiar phrases as 'the search for extra-terrestrial intelligence'. (I have myself used the word 'intelligent' in this strong sense at several points in this book. In Chapter 1, for example, I said that Kant's diagnosis of the failure of human beings to produce a science of metaphysics would apply equally to "intelligent dolphins.") In this phrase, 'intelligence' means exactly what I will mean by 'rationality': anyone who said there was intelligent life

elsewhere in the universe would be taken to mean there were somewhere beings who shared with us mental capacities that the most "intelligent" apes do not share with us.[2]

And what is rationality? Let us begin to try to answer this question by considering another question, a question asked by the philosopher Ludwig Wittgenstein: "We say that a dog is afraid his master will beat him; but not, he is afraid his master will beat him to-morrow. Why not?" The beginning of the answer to this question is that the idea expressed by the word 'tomorrow' is wholly foreign to the mental world of the dog. If the dog can be said to have ideas at all, the ideas that constitute the content of its thought at any moment are ideas of things it is then aware of or of things that might well be immediate consequences of the operations of the things it is then aware of (such as an imminent beating). This point is often put by saying that dogs—and all other non-human terrestrial animals—are "incapable of abstract thought." This idea (applied to a primitive species of our genus—a species more properly called *Homo erectus erectus*) is well expressed in a bit of verse by W. V. Quine:

> *The unrefined and sluggish mind*
> *Of Homo javanensis*
> *Could only treat of things concrete*
> *And present to the senses.*

One might, however, wonder whether dogs and other beasts—other non-human terrestrial animals—are not capable of a *little* abstract thought. After all, "being beaten by one's master" is a sort of abstraction, a universal that has been abstracted from various concrete situations and could have any number of instances. A dog that fears being beaten by its master would seem to fear that something that has happened before will happen again. And it does not fear the occurrence of an exact duplicate of some earlier event; it fears the occurrence of an event that will be the same as a certain earlier event *in a certain respect*: however the feared event may differ from the earlier event, it will be like the earlier event in being a beating by the dog's master. As to the matter of "present to the senses," it suffices to point out that a feared beating that has not yet happened is *not* present to the senses. (It may of course be that it is simply not true that dogs ever fear being beaten, or not in the same sense as that in which human beings fear being beaten. It may be that we use words like these to describe the mental states of dogs simply because we have no others. Perhaps our use of these words is an example of our tendency to anthropomorphism—as when we say, "The sun is trying to come out" or "The car doesn't want to start." But I shall assume our simple, everyday descriptions of the beliefs, hopes, and fears of dogs and other beasts can be literally correct.)

Rationality, then, does not consist simply in the capacity for abstract thought. It consists in the capacity for a certain *kind* of abstract thought. A rational being is a being that can do the following:

> It can represent to itself complex states of affairs, including non-actual states of affairs, that are strikingly remote from its present sense-perceptions. (For example: Jane's coming to visit a week from next Thursday; someone's ordering the second-cheapest item on the menu; the government's preventing a recurrence of bubonic plague by finding a new way to dispose of the refuse that feeds the rats that carry the fleas that are infected with the bacterium that causes the plague.) It can believe that certain states of affairs are actual and that others are non-actual. It can desire that certain states of affairs be actual and others non-actual. It can contemplate states of affairs without raising the question whether they are actual or non-actual. ("I'm trying to imagine what our life will be like if we really go ahead and have a child.") It can be aware of logical and causal relations between states of affairs. It can sort states of affairs into the categories "probable" and "improbable." It can assign relative values to states of affairs. ("I'm sorry I embarrassed you. I didn't *want* to, you know. But I thought that would be preferable to telling an outright lie.") It can devise plans of action that draw on its beliefs about which states of affairs are actual and non-actual and probable and improbable and about the logical and causal relations that hold among both actual and non-actual states of affairs, in order to attempt to cause states of affairs it values to become actual. It is capable of recognizing other beings as having all these capacities, and it is capable of communicating to those that do facts and orders and questions related to the states of affairs it represents to itself and to its beliefs and desires and values in respect of those states of affairs. A rational being, therefore, is a being capable of making statements and giving orders and asking questions; this implies that, in itself and independently of any such communication, it "has" something to make statements and give orders and ask questions about.

This is rationality. (Or, if you like, rationality is *at least* this. I do not deny that rationality may be this and more.) Rationality marks a great divide, a discontinuity between humanity and the beasts. It is wrong to suppose that there is something apes and elephants and beavers have a little of, and we have more of, and that as a consequence, we are rational and they are not.[3]

It is not that we are "more intelligent" than, say, apes, and that that is why we are rational and apes are not—as Alice is able to solve word-analogy problems and spatial-relation problems faster than Alfred because she is more intelligent. (Whatever that means. There. That was a relief. Whenever I write the words 'more intelligent' I feel a very strong urge to add the words 'whatever that means'.) We

may indeed be more intelligent than apes; indeed I suppose we are. But if so, that is not why we are rational and apes are not. If there is a connection, it goes the other way: we are more intelligent than apes because we are rational and therefore have more use for intelligence—for intelligence, if it is anything, is the ability to manipulate mental representations of states of affairs in various useful ways, and we have a lot more, and a lot more complex, representations to manipulate than apes do. To suppose we were rational and apes weren't because we were more intelligent than apes would be like supposing bats could fly and mice couldn't because bats were more "physically agile" than mice. (Bats probably do have greater physical agility than mice—whatever that means. They need greater physical agility because they can fly and mice can't.) Human beings who are of subnormal intelligence owing to injuries or genetic defects do not have minds at all like the minds of apes, any more than apes of subnormal intelligence have minds like the minds of elephants or beavers. Rather, they have human minds that are of diminished capacity in respect of dealing with the demands of life in a human community.

We will consider four questions about rational beings:

- What rational beings are there, and why do they exist?
- What is the place of rational beings in the World?
- What is the nature of rational beings?
- What are the powers of rational beings?

Notes

1. Many people shy away from language like this these days because they believe its use implies that human beings have the right to hunt non-human animals for sport or use them in medical experiments or do just about anything else to them that might occur to us. And many people are opposed not only to engaging in wanton cruelty to animals, but also to eating their flesh and even to using them as sources of wool and milk. It is therefore natural that they should object to language implying that human beings have the right to use their fellow animals in any way they like. But the term 'rational being' has no such implication. One might as well say that to distinguish between animate and inanimate objects is to imply that I, being a living being, have the right to smash Michelangelo's *Pietà* with a hammer. If I am considering a course of action that will affect the welfare of both human beings and dolphins, the fact that human beings are rational animals and the fact that dolphins are not rational animals will quite possibly be *relevant* to the question of the morality of the proposed course of action. But these two facts by themselves could not settle the question.

2. Science-fiction writers have taken to using the word 'sentient' to express the idea I express by 'rational'. But 'sentient' means 'capable of sensation and feeling': dogs and cats are sentient beings.

3. It is wrong but apparently very natural. I once attended a lecture by a specialist in "artificial intelligence" about the enormous difficulties facing anyone who wants to program a computer to be able to talk (like "Hal 9000" in *2001: A Space Odyssey*). A member of the audience asked afterward, in genuine puzzlement, "But why don't you just make the computer very intelligent; if it's intelligent enough, won't it be able to learn to talk?" He was thinking of intelligence and the ability to talk on an "automotive" model: a thing's "intelligence" and its having the ability to talk are related in the way a car's engine power and its having the capacity to move—let's say—as fast as a running cheetah are related: if your car is slower than a running cheetah, just keep increasing the power of its engine, and you will eventually reach a point at which it will be able to match the speed of the cheetah.

10

THE NATURE OF RATIONAL BEINGS

Dualism and Physicalism

Since we know of no rational beings besides ourselves, we shall be able to discuss the problem of the nature of rational beings only in relation to ourselves. We have already said something about the nature of rational beings in one sense of 'nature': we have set out the defining characteristics of rationality. Our question will be this: What is it about human beings that enables them to be rational? Perhaps we can best understand what is meant by this question by drawing an analogy with a question about an everyday physical concept like liquidity. We may know that a "liquid" is a stuff that changes its shape to fit the shape of the container in which it is placed but retains a particular volume throughout all changes of shape. But this does not tell us what it is about water (that is, the chemical compound whose molecules are formed from two hydrogen atoms and an oxygen atom) that accounts for the fact that it is a liquid at temperatures and pressures at which table salt is a solid and carbon dioxide a gas. Explanations of this fact are available. (They appeal principally to the forces that operate between H_2O molecules and the way in which these forces are determined by the properties of hydrogen and oxygen atoms and their arrangement in the H_2O molecule.) We want to find an analogous explanation of the way in which rationality is "realized" in human beings (analogous, that is, to the way in which liquidity is realized in water): we want to know what "underlying" features of human beings enable them to have the properties listed in the abstract definition of rationality.

The short answer to the question, What is it about human beings that enables them to be rational? is, No one knows. The rationality that is, as far as we know,

unique to human beings is a mystery, as is the conscious experience human beings share with many other animals. The two questions 'How is rationality realized?' and 'How is conscious experience realized?' are generally viewed by philosophers as belonging more to the part of philosophy called "the philosophy of mind" than to metaphysics. Or at least this is true when these questions are considered in their entirety. But there is a question that could be thought of as a part of these questions (an answer to it would be a part of the answers to them) that is pretty clearly within the domain of metaphysics. We shall devote this chapter and the following chapter to this question.

The question we shall be addressing is rather hard to state if we want to state it in a way that does not favor one answer to it over other possible answers. We might try this: What *kind* of thing are we human beings? But this formulation is too abstract to convey much. It often happens in philosophy that philosophers pose a question and suggest various answers to it, and that the answers are clearer than the question. The present case is one of them. One way to deal with such a difficulty is to let the answers define the question: it is the question to which those statements are possible answers. Let us try that strategy.

The possible answers to the question we are trying to understand (at least the possible answers that are taken at all seriously today) are all forms of either *dualism* or *physicalism*. The first step in trying to understand our question is to understand these terms.

Suppose that by a "physical thing" we mean an individual thing made entirely of those things whose nature physics investigates. If current physics is correct, all the objects of our sensory experience—pieces of chalk, beetles, stars, and everything else we can touch or see—are made entirely of three kinds of elementary particles: up-quarks, down-quarks, and electrons (plus a few kinds of particles, such as photons, whose exchange by quarks and electrons enables the quarks and electrons to interact). It is an interesting technical question what we mean by 'made entirely of', but let us suppose we have an adequate intuitive understanding of this phrase. (Here is an example to aid our intuitions: A sand castle is made entirely of grains of sand—provided the child who built it did not incorporate into its structure a twig or lollipop stick or anything else not made of sand.) Thus, by the terms of our definition, all the objects of our sensory experience are physical things.

If an individual thing neither is itself a physical thing nor has any physical things as parts, we shall call it a "non-physical thing." We should note that this definition does not rule out the possibility of individual things that are neither physical things nor non-physical things. An object that had both physical things and non-physical things as parts would be neither a physical thing nor a non-physical thing. We could call such an object an "amalgam." I shall have nothing to say about amalgams, apart from a few brief remarks in the notes. When I talk of

things that are "not physical," my remarks are meant to apply only to non-physical things and not to amalgams, even though amalgams are, strictly speaking, not physical things. (And my remarks apply only to individual things. Universals are not non-physical things in the sense I am giving the term, despite the fact that universals are not physical things.)

In addition to the concept of a physical thing, it will occasionally be useful to have the concept of a physical *property*: we shall understand a physical property to be a property that can be possessed by a physical thing and *only* by a physical thing.

Since we can see and touch human beings, and since we are human beings, it might be thought to follow from our definition of a physical thing that we are physical things. But let us make some distinctions. Let us say that a *human organism* is that which a biologist would classify as a member of the species *Homo sapiens*. And let us say that a *human person* is that which we refer to when we use the first-person-singular pronoun ('I', 'me', 'moi', 'ego', 'ich', . . .). When I have used the words 'human being' in this and earlier chapters, I have been assuming that human persons and human organisms are one and the same. To call *x* a human being is to call *x* a human person, but with the understanding or implication that *x* is a human organism, a rational animal. (Or this, at least, is what I take 'human being' to mean. Perhaps there are those who would dispute this definition.) But the thesis that human persons and human organisms are one and the same is controversial.

If human persons and human organisms are one and the same, then, since human organisms are obviously physical things, it follows that human persons are physical things. The thesis that human persons are physical things is called *physicalism*. (This word is also used as a name for the stronger thesis that *all* individual things are physical things. And the stronger and weaker senses of the word tend not to be carefully distinguished, owing to the fact that most philosophers who believe that human persons are physical things also believe that all individual things are physical things. I shall use 'physicalism' only for the thesis that human persons are physical things.[1])

The thesis that human persons are non-physical things is called *dualism*. (More exactly, the thesis that there are both physical and non-physical things and that human persons are among the non-physical things is called dualism. Some idealists perhaps hold that there are only non-physical things, persons among them; such idealists are not dualists.) This word comes from the Latin word for 'two'. The dualist believes that human persons have a "dual" nature. The person is, strictly speaking, a non-physical thing, but it is very intimately associated with a certain physical thing, a human organism, which is called the person's *body*. The body, not the person, is the thing a biologist would classify as a member of the species *Homo sapiens*. The dualist will concede that we frequently make assertions by which we appear to ascribe physical properties to human persons, assertions like, "John

weighs 90 kilograms" or "Alice is 165 centimeters tall." But according to the dualist, it is not strictly true that John weighs 90 kilograms or has any other weight, and it is not strictly true that Alice is 165 centimeters tall or has any other height. John and Alice, rather, possess such properties only vicariously; strictly speaking, it is not they but their bodies that have weights and heights. This does not mean that there is anything wrong with saying "John weighs 90 kilograms" in ordinary contexts; this statement is to be understood as a kind of shorthand expression of the assertion that John's body weighs 90 kilograms, just as Alice's statement "I'm carrying 1,400 tons of pig iron" is a shorthand expression of the assertion that the ship of which she is the cargo officer is carrying 1,400 tons of pig iron. A "dualistic" analysis of the ordinary statement "John weighs more than he likes" well illustrates what is meant by saying that, according to the dualist, human persons have a "dual nature." Nothing, according to the dualist, could literally weigh more than it liked. Rather, the dualist holds, it is John, the non-physical person, who does the disliking, and it is his body, the physical organism, that has the weight that is the object of the dislike.

What is the "intimate association" that holds between the person and the person's body? Dualists have answered this question in more than one way. The most obvious answer, and the one that commands the widest allegiance among dualists, is contained in a theory called "dualistic interactionism." In order to set out the content of this theory, let us look at a typical human person and see what dualistic interactionism says about the relations that have to hold between a person and an organism for that organism to be that person's body. Let us consider one Jane Tyler, the author of the well-regarded novel *The Sinews of Thy Heart*, whom we may suppose to be a typical human person. And let us consider the following words and phrases:

- 'Jane Tyler'
- 'the author of *The Sinews of Thy Heart*'
- 'I' (spoken by Jane Tyler)
- 'you' (spoken by someone addressing Jane Tyler)
- 'she' (spoken by someone relating an anecdote about Jane Tyler)
- 'that woman over there' (spoken by someone calling someone's attention to Jane Tyler)
- 'Jane Tyler's mind'
- 'Jane Tyler's soul'

According to the dualist, when these phrases are spoken in the indicated contexts, they denote or name or stand for or refer to the same thing, a non-physical thing, a thing not composed of elementary particles and not observable by the senses, a

thing without weight or mass (gravity and inertia are concepts that apply only to physical things) and having no position in space—at least it is hard to see how a non-physical thing could have a position in space, although Saint Thomas Aquinas believed that angels were non-physical things that had positions in space. (The dualist will probably also want to say that this thing has no parts: as metaphysicians say, it is a *simple*. But in principle, one could be a dualist and hold that a human person had parts, provided they were all non-physical parts.)

In addition to Jane Tyler there is Jane Tyler's body, a physical thing, a living human organism. Our question is: What is it that makes one particular human organism *Jane Tyler's* body and not some other person's body—or no one's body at all? Dualistic interactionism tells us that this particular organism is Jane Tyler's body because of a certain two-way causal connection that holds between Jane—let us get on familiar terms with her—and that organism. A certain organism is Jane's body because she affects it and it affects her. But we must be more specific than this, because cause-and-effect relations can hold between any human person and any human organism.

There is, interactionists maintain, a very special way in which Jane can affect the one particular human organism that is her body: she can cause changes in it without causing changes in any other organism (other than its own parts; multicellular organisms have cells, which are themselves organisms, as parts). And there is a very special way in which one particular organism can affect her: it can cause changes in her without causing changes in any organism besides itself (and its own parts).

Let us look at an example. Suppose Jane begins to whistle. In doing this she causes changes in a certain organism (electrical currents flow along very specific neural pathways in the organism, its lips assume a specific configuration, and many other changes occur in it). And it may be that in beginning to whistle, she causes changes in no organism but this one and some of its constituent cells. Now *I* can also do things that will cause changes in that organism; I can, for example, open a window on a freezing day and cause it to begin to shiver. But I can do this only by causing changes in another, wholly distinct, organism, *my* body.

Now let us look at an example of the special way in which changes in the organism that is Jane's body can cause changes in Jane the person. Suppose Jane steps on a tack. The resulting puncture wound in her foot will cause *her* to be in pain. (Being in pain would seem clearly to be a property of Jane the person. Being in pain—having the *sensation* we call "pain"—is a property of an organism only if the organism, or some part of it, *is* a person.) It is true that changes in other organisms than Jane's body can cause changes in Jane. If I step on a tack, the resulting puncture wound in my foot may cause her to feel concern (and feeling concern is a property of the person). But a change in my body can cause a change in Jane only by causing a change in another organism, *her* body, that is not a part of my body.

Dualistic interactionism, then, consists of two theses: dualism, the thesis that there are human persons and human organisms and that no human person is a human organism (or any other physical thing), and interactionism, the thesis that each human person (at any rate, each living human person) has a body, a unique human organism to which it is bound "directly" by mutual causal interaction. The two most important dualists in the history of metaphysics, Plato and Descartes, were interactionists. Other dualists, however, have rejected interactionism, generally because of the physical or metaphysical difficulties raised by the thesis that a non-physical thing (a thing having no physical properties like mass or electrical charge) could affect a physical thing. Descartes's follower Nicholas Malebranche, for example, held that when a person "wills" or "tries" or "sets out" to whistle, God effects appropriate changes in a certain human organism. Similarly, he held that when a human organism is punctured by a tack, God causes a certain person to experience appropriate sensations of pain. This theory is called "occasionalism," since it holds that changes in the person are never the *causes* of changes in an organism but are only the "occasions" of changes in an organism; in the same way, changes in an organism are never causes of, but only occasions of, changes in a person.

A second dualistic alternative to interactionism is "epiphenomenalism" (from a Greek word meaning 'by-product'). According to this theory, changes in a human person can be caused "directly" by changes in a particular human organism, but changes in the person never cause changes in that organism. Each change in the organism is caused by prior changes in the organism or in its immediate physical environment, and these physical events also sometimes cause changes in the person—but there is no "feedback" from the person to the organism: the non-physical events that are changes in the person never have physical effects. Persons are thus related to their bodies as billows of smoke are to the fires from which they issue: persons exist and are non-physical things, but they are mere by-products of the physical activity going on in certain organisms. (Or this is one way to understand epiphenomenalism. Epiphenomenalists have not generally expressed themselves very clearly. It is possible that at least some epiphenomenalists want to say that the person *is* the organism and that it is people's *sensations and thoughts* that are the by-products of the events going on in the organism. Other epiphenomenalists write in such a way as to suggest that persons are not individual things at all but are mere collections of the thoughts and sensations generated by "their" organisms. I can make nothing of either of these ideas.) It is a consequence of this theory that our belief that we can influence the motions of our bodies is an illusion. The illusion is itself, according to epiphenomenalism, a by-product of the physical activity of the body.

There are several other dualistic theories of the nature of the person-body relation, but we shall not discuss them. Nor shall we further discuss occasionalism and epiphenomenalism.

We should take note of one other point about dualistic interactionism: it does not obviously follow from dualistic interactionism that the non-physical human person can exist without being in interaction with a human body. Some argument would be required to establish that a dualistic interactionist should believe a human person could exist without a body. Plato believed that the soul—that is, the person—would "automatically" continue to exist when the body it was associated with died. And he did have an argument for this thesis: that the soul is a metaphysical simple, and that a thing can cease to exist only by "coming apart," by being resolved into its elements; a simple, a thing without parts, must therefore be imperishable. This argument, however, is not particularly convincing. For example, the premise that a thing can cease to exist only by coming apart deserves further discussion. One might cite the fact that current physics treats electrons and various other particles as having no parts; yet an electron can be "annihilated" by a collision with a positron. But we shall not pursue this subject. We shall not try to discover whether Plato's argument is ultimately defensible or whether there might be other interesting arguments for the same conclusion.

The physicalist, who holds that the human person just *is* the human organism (or some part of it), does not face the problem of explaining the relation between person and organism.[2] Since for the physicalist the person and the organism (or a part of the organism) are identical, a change in the person is a change in the organism. And since the organism is a physical thing, and a physical thing is made entirely of quarks and electrons, it would seem that any change in a human person must be a change in the physical properties of the person: a change in the properties of the quarks and electrons that make the person up, or else a change in the way the quarks and electrons that make the person up are related to one another. Such a change—a change in the physical properties of a thing—we may call a physical change; examples of physical changes would be *receiving a puncture wound in the foot*, *undergoing a sudden rise in body temperature*, and *having a brain in which electrical currents suddenly begin to flow in such-and-such a way*.[3] If a human person is a physical thing, any change whatever in a human person must be a physical change. If, for example, Tim becomes elated because of some news contained in a letter he has just received, this change in Tim, his becoming elated, must be the very same thing (or perhaps we should say the very same event) as some physical change.[4]

If it is indeed true that Tim's becoming elated is the very same thing as some physical change, then, given what we know about human physiology, it is

presumably the same event as some event involving some of the particles that make up Tim's brain—no doubt a change in the way in which electrical currents flow in Tim's brain. Thus, if physicalism is correct about the nature of persons, all those changes in a person we unreflectively call "mental" or "psychological"—whatever, exactly, these terms may mean—are physical changes in the person (and presumably changes in the person's cerebral cortex, the part of the brain associated with conscious mental activity). The thesis that mental changes (in human persons at least) just *are* certain physical changes is called the "identity theory." The identity theory is not quite the same thing as physicalism. Physicalism (the theory that human persons are physical things) entails the identity theory (that mental changes in human persons are identical with certain physical changes) only on the assumption that mental changes in human persons really exist. And there are philosophers and psychologists who deny the existence of the mental (mental changes and mental states) altogether. We shall not discuss the views of these philosophers and psychologists, who subscribe to theories with names like "behaviorism" and "eliminative physicalism." We shall take the reality of the mental for granted, as do most philosophers and psychologists and, indeed, most physicalists. (Because most physicalists take the reality of the mental for granted, it is safe to say that most physicalists subscribe to the identity theory.)

The two most important theories about the nature of the only rational beings whose existence is uncontroversial (ourselves) are, therefore, dualistic interactionism and physicalism. What can be said for and against each of these theories? Can either be shown to be superior to the other?[5]

We shall begin our attempt to answer these questions by examining some arguments for dualism. (We shall not concern ourselves with defending dualistic *interactionism*; we shall take it for granted that interactionism is the most plausible form of dualism and shall investigate the question, What can be said in defense of dualism?) Arguments for dualism have this general form: you and I and other human persons are not human organisms or any other physical things, because we have properties that could not belong to a physical thing. (It is obviously a valid general principle of reasoning that a thing x and a thing y cannot be identical, cannot be one and the same thing, if x has a property or feature or characteristic that y lacks.) There are many such arguments. We shall consider five of them. The first argument we shall examine is commonly ascribed to Descartes. (Some commentators find this argument in his *Meditations on First Philosophy*, others in his *Principles of Philosophy*. The passages in both books in which the argument can supposedly be found are, it must be confessed, rather obscure. But the argument is an interesting argument whether or not it is Descartes's. Without pretending to have settled any textual point, I will, simply as a matter of literary convenience, ascribe the argument to Descartes.)

This is Descartes's argument: I can conceive of my body's not existing—indeed, I can conceive of there being no physical world at all—but I cannot conceive of *my* not existing; I am therefore not my body.

When Descartes says I can conceive of my body's not existing, he is not advancing the thesis that I can form a conception of the way things would have been if my body had not existed (no doubt I can, but that I can is not his thesis); he is advancing the stronger thesis that it is possible for me to conceive of the following: things being *just as they seem to me to be* and yet there being no such thing as my body. To conceive of this, I could imagine that there exists some powerful spirit (the "evil genius" we met in Chapter 3) who has decided to deceive me about the existence of a world of physical things: there are no physical things, but the spirit deceitfully "feeds" me a series of sense impressions like the series of sense impressions I should be experiencing if I were perceiving a world of physical things.

And when Descartes says that I cannot conceive of *my* not existing, he is not saying that I cannot form a conception of the way things would have been if I had not existed (that would be false; I can conceive of that); he is saying rather that I cannot conceive of the following: *things being just as they seem to me to be* and yet there being no such thing as myself. In other words, Descartes holds that, however absurd it may seem, the hypothesis that I exist and no physical thing exists (which of course implies that I do not have a body) is an hypothesis it is *possible* for me to entertain; but the hypothesis that I do not exist is not simply an hypothesis that it is impossible for me to entertain without absurdity: it is an hypothesis it is impossible for me to entertain—impossible *full stop*, impossible *period*. It is remotely possible that my conviction that there are physical things, including my own body, is an illusion. It is not even *remotely* possible that it is an illusion of mine that I exist. Not an illusion of *mine*: if I am "there" to have the illusion, I must exist.

The argument, then, is that my body has the following property:

can be conceived by me not to exist,

as does every other physical thing. But *I* do not have that property. Therefore, I am not identical with my body—nor am I identical with any other physical thing.

The trouble with this argument is that it proves too much. I can obviously make *some* statements of the form 'I am (identical with) . . . ' (where the blank is to be filled in by something other than 'I' or 'me' or 'myself') and thereby say something true, but an argument having the same form as Descartes's argument can be used to refute any such statement. Let us look at an example. The statement

I am the author of *An Essay on Free Will*

is true; that is, if I were to speak these words, I should say something true, for there is a book of that title, and I am its sole author. But suppose I were to reason as follows:

> I can conceive of there being no such thing as the author of *An Essay on Free Will*. That is, I can conceive of things being just as they seem to me to be and there being no such thing as the author of *An Essay on Free Will*. The easiest way would be for me to suppose that there is no such book: my apparent memories of having written and published such a book are fantastic delusions. But I cannot conceive of there being no such thing as myself. Therefore, the author of *An Essay on Free Will* has the property "can be conceived by me not to exist" and *I* do not have that property. Therefore, I am not the author of *An Essay on Free Will*.

Since this argument starts from true premises and yet has a false conclusion, it must contain some error of logic. Most philosophers would agree that the error is this: the words 'can be conceived by me not to exist' do not name or express a property, but the argument treats them as if they did. If these words did name or express a property, we ought to be able to take a sentence like 'The author of *An Essay on Free Will* can be conceived by me not to exist' and substitute for 'the author of *An Essay on Free Will*' any word or phrase that denotes (designates, refers to, is a name for) the same thing and get a sentence that is true if the original sentence is true.

But this is not what in fact happens. The word 'I' denotes (when *I* use it) the same thing as 'the author of *An Essay on Free Will*'; but 'The author of *An Essay on Free Will* can be conceived by me not to exist' is true, and 'I can be conceived by me not to exist' is false. Let us compare 'can be conceived by me not to exist' with some phrase that really does name a property—say, 'was born during the Second World War'. The author of *An Essay on Free Will* was born during the Second World War (take my word for it). The word 'I', when I speak it, and the words 'the author of *An Essay on Free Will*' are two names for the same thing. The appropriate substitution produces the sentence 'I was born during the Second World War'. Is it true that I was born during the Second World War? Well, of course it is. It has to be, given that the author of *An Essay on Free Will* was born during the Second World War and that I am the author of *An Essay on Free Will*.

If a phrase that looks as if it named a property (like 'can be conceived by me not to exist') does not obey this simple substitution rule, then contrary to appearance, it does not name a property. Therefore, 'can be conceived by me not to exist' does not name a property. And therefore, Descartes's attempt to prove that persons are not physical things contains an error. There is nothing wrong with the principle of reasoning 'If *x* has a property *y* lacks, then *x* is not identical with y', but Descartes

misapplied this valid principle as a result of his treating 'can be conceived by me not to exist' as a name of a property.

We now turn to our second argument for dualism, a very popular one:

> Physical things are incapable of thought and sensation. But human persons are capable of thought and sensation. Therefore, human persons are not physical things.

But why should we believe that physical things are incapable of thought and sensation? I am willing to grant that if we try seriously and in detail to imagine a physical thing having thoughts and sensations, we can find this notion—the notion of a physical thing having thoughts and sensations—very puzzling. There is a famous passage in Leibniz's *Monadology* that very clearly brings out the puzzling aspects of this notion:

> Furthermore, we must admit that *perception*, and whatever depends on it, *cannot be explained on mechanical principles*, i.e. by shapes and movements. If we pretend that there is a machine whose structure makes it think, sense and have perception, then we can conceive it enlarged, but keeping to the same proportions, so that we might go inside it as into a mill. Suppose that we do: then if we inspect the interior we shall find there nothing but parts which push one another, and never anything which could explain a perception. Thus, perception must be sought in simple substance, not in what is composite or in machines.[6]

To take a more modern example, suppose someone were to claim to have programmed a computer so that it could think (in a sense that implies conscious experience and self-awareness) or to have constructed a thinking robot. If the computer or robot were enlarged so that people could walk about inside it, a party of tourists being led through the vast machine would see nothing but physical things interacting physically. And this would be no illusion. It's not as if the thought and conscious experience were hidden away in some part of the machine off limits to visitors.

But then where *are* the thoughts and the experience? Where *could* they be? How could the mere physical interaction of bits of metal and plastic and silicon "add up to" thoughts and experience? It is important to realize that this point has nothing to do with the specific kinds of physical material a computer or robot would be likely to be made of. The point has to do only with the fact that the materials are *physical*. The point would be unchanged if we imagined a party of tourists being conducted through *ourselves* (or our bodies), as in Isaac Asimov's interesting science-fiction novel *Fantastic Voyage* (or the unspeakably

silly movie of the same title). If we could be greatly reduced in size and go inside a functioning human brain and have a look round, we should see no thoughts or experience, not even if we saw everything there was to see. If God looks inside a human brain, even He sees nothing but unthinking physical things like neurons and Nissl granules and amino-acid molecules and electrons in continuous mutual physical interaction. Where, then, are the thoughts? Where are the sudden feelings of elation or despair? Where are the sensations of heat and pain and pressure and color? The answer is, obviously, that they are elsewhere. And that "elsewhere" must be a place that is receptive to the presence of such things, a place where they *could* exist. They must exist in a non-physical thing. (If we like, we can say that they must exist in a non-physical thing that is *mental*: a mind or a soul. But unless we can say something useful about what we mean by 'mental thing' or 'mind' or 'soul', to say this would be to say no more than that they must exist in a non-physical thing.)

Various physicalists—who must of course believe that physical things are capable of thought and sensation—will reply to this argument in various ways. What follows is my own reply. Some physicalists would reject some parts of it.

Let us begin with the question, *Where* are the thoughts and sensations? The answer is that since these things are changes in the cerebral cortex, they are all around you (you who have in imagination been reduced in size and are physically inside someone's brain). It does not follow from this that you *see* them, since they may involve the whole cerebral cortex or the whole brain or widely scattered parts of the brain: it may be that you cannot see them for the same reason you cannot see the event called 'the election' on election day. But let us suppose for the sake of argument that these events are sufficiently localized that you can see them. (Or some aspects of them: a human being cannot see every aspect of any event. You can see the street lamps come on in your neighborhood, but you cannot see the flow of electrons that is an indispensable component of this event.) Of course these events do not *look* to you like mental events, but then what would you expect a mental event to look like? ("Well, something like the way mental changes in *myself* look to *me*, as when I experience a sharp pain in my left shoulder or a thrill of fear or an intellectual insight." But that's what it's like to experience *having* or *being the subject of* a mental change. That's what a mental change *in you* "looks like" to you. What would you expect mental changes in someone *else* to look like to you?) And anyway, a change may be of a certain type without its being evident that it is of that type. Suppose a computer has been programmed to compute the orbit of a certain satellite. Suppose the computer were greatly enlarged and that you went inside it, "as into a mill." You would not see any orbital computations going on—or at least you would not see anything that "looked like" orbital computations. (What would you expect orbital computations to look like?) The Leibnizian thought-experiment,

therefore, should cause the physicalist no unease. Things inside the brain look just the way they would look if physicalism were correct.

Many physicalists would think that this was a sufficient reply to the charge that the notion of a physical thing that thinks is mysterious. I cannot agree with them. I do not deny that everything said in the preceding paragraph is correct, as far as it goes. Nevertheless, it seems to me that the notion of a physical thing that thinks is a mysterious notion, and that Leibniz's thought-experiment brings out this mystery very effectively. We must remember, however, that our present question is not whether the physicalist is faced with a mystery; our question is whether dualism is to be preferred to physicalism. If thinking is a mystery for the physicalist, this fact will be relevant to our question only if it can be shown that the dualist is not confronted with the same mystery or some corresponding mystery.

And, I believe, the dualist is. For it is thinking itself that is the source of the mystery of a thinking physical thing. The notion of a non-physical thing that thinks is, I would argue, equally mysterious. How any sort of thing could think is a mystery. It is just that it is a bit easier to see that thinking is a mystery when we suppose that the thing that does the thinking is physical, for we can form mental images of the operations of a physical thing, and we can see that the physical interactions represented in these images—the only interactions that *can* be represented in these images—have no connection with thought or sensation, or none we are able to imagine, conceive, or articulate. The only reason we do not readily find the notion of a non-physical thing that thinks equally mysterious is that we have no clear procedure for forming mental images of non-physical things. Still, we are not wholly without resources for constructing mental images of non-physical things. (No doubt most of us associate some sort of mental image with the doctrine of dualistic interactionism: perhaps a human body with a vague "something" inside or above its head.) Let us see what we can do.

Leibniz, in the passage we have quoted, contends that a thinking thing must be a simple, a thing without parts. Well, let us represent, in our thought, a simple non-physical thing by a dot and a composite non-physical thing by a bunch of dots, perhaps a bunch that is in constant internal motion like a swarm of bees. Might a *composite* non-physical thing "think, sense, and have perception"? It is hard to see how. Consider our proposed mental picture of a composite non-physical thing. If the simples that make up a composite non-physical thing do not think individually, where is the thinking in our picture? How can a bunch of things that do not individually think or sense or have perception add up to something that does think or sense or have perception? How could their causal interaction produce such properties? Note that these questions are exactly parallel to the questions Leibniz's thought-experiment raises about thought and composite *physical* things. The only real difference between the two cases is that a mental image of a

composite physical thing will have reasonably "sharp" constituents drawn from our experience of actual physical things—images of gears and wheels, say—, whereas (an attempt at) a mental image of a composite non-physical thing will be vague and arbitrary (arbitrary because non-physical things necessarily lack visual characteristics; we chose dots because dots come as close to having no characteristics as anything we can picture).

Leibniz would no doubt agree that these reflections show that a composite non-physical thing cannot think. After all, his position is that a thinking thing has to be a simple.[7] But let us look at our proposed mental picture of a (non-physical) simple. It is just a dot. How can we cause it to change in our imagination in such a way that this change will represent its having a series of thoughts and sensations? Change of position (relative to other imagined dots) will be of no help, because that is a relational change, and thought and sensation are supposed to be intrinsic features of thinking, sensing things. Even a dot must have a shape, but when we use dots to represent non-physical simples we do our best not to attend to their shapes, for insofar as we think of a dot as having a shape, we think of it as being composed of smaller regions and thus as composite.

We might think of the dot as changing color, I suppose. Let's try that. Imagine a dot continuously changing its color in some very complex way. Are you imagining something thinking or having sensation? Where are the thought and the sensation in the picture your imagination has created? My point in asking these unanswerable rhetorical questions is not to suggest that a non-physical simple cannot think. (Although I believe that human persons are physical things made of smaller physical things, I believe that God is a non-physical simple, so I should hardly want to suggest that a non-physical simple cannot think.) My point is that nothing could possibly count as a mental image of a thinking thing. Or at least, nothing could count as a mental image that *shows* or *displays* a thing as thinking (except by convention, as, for example, "thought-balloons" in comic strips do, or via the familiar outward and visible signs of human thought, like those displayed by Rodin's *The Thinker*). And, I am suggesting, we need to keep this fact in mind when we consider Leibniz's thought-experiment. It is only the difficulty of conducting a similar thought-experiment for non-physical things that keeps us from seeing that his thought-experiment does not favor dualism over physicalism. Consider this analogy. We are amazed to see a human figure hurtling through the sky like Superman. "It's a woman!" someone shouts. "Why a woman?" we ask. "Well, it's either a man or a woman, and it's impossible for a man to fly." This argument is valid, and there are certainly good reasons for thinking that it's impossible for a man to fly. But there are equally good reasons (the same ones) for thinking that it's impossible for a woman to fly. Therefore, the argument gives us no reason to prefer the hypothesis that the human figure we saw in the sky was a woman to

the hypothesis that it was a man. And this is exactly parallel to what one should say in response to Leibniz's thought-experiment: Since we are unable to imagine a non-physical thing in a way that displays it as thinking, the fact that we are unable to imagine a physical thing in a way that displays it as thinking does not give us a reason to prefer the hypothesis that we human thinkers are non-physical things to the hypothesis that we are physical things.

These points about mental images can be generalized so as to apply to any type of representation. Mental images are representations of how things are or might be, but there are representations of many other kinds, such as schematic diagrams on paper, three-dimensional cardboard models, computer models, and scientific theories. In general, to attempt to explain how an underlying reality generates some phenomenon is to construct a representation of the working of that underlying reality, a representation that in some sense "shows how" the underlying reality generates the phenomenon. (The best scientists seem to be able to "translate" their verbally and mathematically formulated representations of the workings of things into images, which they are able to manipulate mentally in fruitful ways.) Essentially the same considerations as those that show that we are unable to form a mental image that displays the generation of thought and sensation by the workings of some underlying reality (whether the underlying reality involves one thing or many, and whether the things it involves are physical or non-physical) show that we are unable to form *any* sort of representation that displays the generation of thought and sensation by the workings of an underlying reality. Thought and sensation are therefore a mystery—although not necessarily an insoluble one. But since the mystery, soluble or insoluble, is entirely independent of whether the elements in the representation are supposed to represent physical or non-physical things, the mystery of thought and sensation does not favor dualism over physicalism.

Has the dualist any way to respond to this counter-argument? The answer to this question depends, I believe, on what the dualist can tell us about the positive nature of the non-physical thinking things whose existence dualism asserts. If the dualist can say no more about them than that they are non-physical things, dualism gains no advantage over physicalism and perhaps gains the disadvantages of postulating the existence of things of a kind physicalism does not postulate and of having to account for the interaction between these things and physical things. Let us (once more) consider an analogy. Suppose Sir Aaron Oldham, the well-known imaginary seventeenth-century scientist, set out to explain the observed phenomenon of magnetism. Sir Aaron believed that all physical interaction was transmitted by contact between physical things, by "pushes and bumps," and he was therefore unable to believe that magnetism was a wholly physical phenomenon, since it could act across empty space and could act "through" a physical object like a sheet

of glass or paper without affecting the intermediate object in any way. He therefore postulated that associated with each lump of lodestone (the only magnets he knew about) there was a non-physical thing that had the power to cause nearby iron objects to move toward the lodestone. "Should a Lodestone be enlarged," he wrote, "to such a degree that a Man were enabled to pass amongst the corpuscules composing it, as an Earthworm might pass amongst the particles of Soil comprised in my Garden, he would observe nought but corpuscules, whether at rest or in motion, a certain quantity of Motion being on frequent occasion translated from one to another of the same corpuscules by Collision. He would see therein no Action by which the motion of a distant Pin or Nail toward those corpuscules might be effected."

We may imagine—let us shift to the historical present—that one of Sir Aaron's scientific rivals puts forward an alternative theory of magnetism: that there are unknown physical interactions, interactions other than pushes and bumps, that cause pins and nails to move toward lumps of lodestone. It would seem that unless Sir Aaron can say something about the positive nature of the non-physical entities he has postulated—unless he can say something more about them than that they are non-physical—his theory enjoys no advantage over that of his rival. (Unless Sir Aaron and his rival tell us more than they have so far, this is how things stand: each theory ascribes an observed phenomenon to an unknown cause and tells us nothing about that cause that explains how it produces the phenomenon.) And it might be argued that Sir Aaron's theory is burdened by a disadvantage his rival's is free of: it postulates the existence of non-physical things in addition to physical things, and it faces the problem of explaining how the non-physical can interact with the physical.

Can the dualist tell us anything positive about the nature of human persons? Can the dualist say anything more about human persons than that they are *not* physical things? Many dualists think they can. In this they follow Descartes, who held that the essence of a human person was thinking. This would appear to mean that the *only* intrinsic properties a human person has or could have are "mental" properties—that is, properties that imply either thought or sensation (and that the human person is essentially such: no human person could possibly have any intrinsic properties but mental properties). Thus, if Descartes is right, human persons have such properties as *being in pain* and *feeling depressed* and *wondering how to spend Saturday afternoon*; human persons do not and could not have such properties as *being 165 centimeters tall* or *weighing 90 kilograms* or any other intrinsic non-mental property.[8]

A typical physicalist believes that human persons have both mental and non-mental properties. A dualist might believe this also, although the dualist, unlike the physicalist, would have to say that the non-mental properties of the human

person were not physical properties, either—that they were, perhaps, the members of some utterly unknowable class of properties. A dualist of this sort might even hold that our mental properties were related to these "other" properties in the way in which the typical physicalist holds that our mental properties are related to our physical properties: as the typical physicalist thinks that physical properties underlie and determine our mental properties, so the dualist might hold that the "other" properties underlay and determined our mental properties. A dualist *might* hold this, but few if any dualists do, and Descartes certainly does not. Descartes's position is that we are mental "all the way through."

Dualists therefore have available to them an account of the positive nature of the non-physical human person: the human person is a mental thing—loosely speaking, a thing having only mental properties. (At least the dualists have such an account available to them if they can solve the very difficult technical problem raised in note 8. In the sequel, I shall assume that they have somehow solved that problem.) And most if not all dualists accept this account of the positive nature of human persons. They have, therefore, an answer to the charge that they have accounted for the phenomenon of thought and sensation simply by postulating a cause for this phenomenon whose positive nature is entirely unknown.

Does their ability to offer this positive account of the nature of human persons provide a reason for preferring dualism to physicalism? It is, I think, plausible to argue that in offering this positive account they have done essentially what Sir Aaron Oldham would have done if he had attempted to give an account of the positive nature of the non-physical things associated with lumps of lodestone by saying that these things had "magnetic" properties and no others. That would not really be an "account" at all, because the words 'magnetic property' could mean nothing but 'power to produce the observed phenomenon of magnetism'. We should have no "hold" on what a magnetic property was except through its observed effects, the very things we want to explain. The dualist who maintains that we are things that have only mental properties is simply asserting the existence of things that manifest the phenomenon to be explained (thought and feeling) and which have no properties besides that of manifesting the phenomenon. It is important to stress that this argument does not have the least tendency to show that dualism is wrong. For all we have said so far (note 8 aside), there might well be things that had only mental properties. The argument is not designed to show that dualism is wrong, but only that dualism enjoys no advantage over physicalism as regards the mystery of thought and sensation.

The dualist who asserts that thoughts and sensations occur as changes in a thing all of whose properties are mental has done no more to address the mystery of thought and sensation than has the physicalist who asserts that thoughts and sensations occur as changes in a physical thing. It is true that no one has any account

of how thoughts and sensations could be features of physical organisms. In fact, no one can say what an account of this would look like, even in broadest outline. But then no one has any account of how there could be a thing that had only mental properties, and no one can say what an account of this would look like, even in broadest outline.

We now turn to a third argument for the conclusion that one is not the same thing as one's body. (That is, for the conclusion that one is not the same thing as the human organism one can bring about changes in without bringing about changes in any other multicellular organism.) This argument proceeds from the observation that we do not seem to ourselves to occupy the same regions of space as our bodies. The twentieth-century English philosopher G. E. Moore formulated this observation in a strikingly simple phrase: "I am closer to my hands than I am to my feet." (Think about it. Look at your hands and your feet at the same time. Your feet are farther away, aren't they?) But my body is obviously not closer to my hands than to my feet—to say it was would be like saying Europe was closer to Belgium than to Italy.

The first thing to note about this argument is that, unlike the two arguments we have so far examined, it does not even claim to prove (in my case) that I am not a physical thing. It claims to prove only that I am not a *certain* physical thing: my body. Even if the conclusion of the argument is true, I might be my brain or my left cerebral hemisphere or my cerebral cortex, for those things are all closer to my hands than to my feet. And of course the argument has the same limitation when it is applied to you or to any other human person. One might even maintain that it is inconsistent with dualism to suppose that I am closer to my hands than to my feet. I can be closer to my hands than to my feet only if I have a position in space, and as we have remarked, it is hard to see how a non-physical thing could have a position in space.

The argument is, however, unconvincing even as an argument for the conclusion that one is not one's body. There may be a sense in which it seems to me that I am closer to my hands than to my feet, but this appearance might be mere appearance and not reality. Our sense organs—leaving aside the skin, our organ of touch—cluster around the brain. Is it not plausible to suppose that one might seem to oneself to be located at or near the place where one's sense-organs cluster? We seem to ourselves to be at the center of the environment our senses reveal to us, and if our sense-organs cluster around some small region, that region will seem to be at the center of our "subjective world." In fact, it is plausible to suppose that sighted persons would seem to themselves to be approximately where their eyes were, even if their ears and other sense-organs were moved to their elbows and ankles, for sighted people construct their internal model of their immediate environment mainly on the basis of visual data. (Consider Helen Keller, who was

blind and deaf from very shortly after her birth. Her model of her immediate surroundings was based almost entirely on tactile data, the data of touch. Would she have felt it natural to say she was closer to her hands than to her feet? Well, perhaps she would have, given the central role her hands played in her knowledge of her immediate environment. But perhaps she would also have felt it natural to say she was closer to her arms than to her head. One can imagine her touching her arms and saying, "My arms are right here . . . ," and then reaching up to touch her head and saying, " . . . but my head is way up here.")

Our fourth argument for the conclusion that we are not physical things proceeds from the premise that whether or not there are other rational beings in the cosmos, there certainly could be: there is nothing intrinsically impossible in the notion. And there is nothing intrinsically impossible in the notion that such beings might be physically very different from us. Therefore, it is intrinsically possible for there to be beings that have thoughts and feelings very much like ours, even though they are *radically* different from us in their anatomy and physiology. Imagine a science-fiction story in which there are beings, the Scorpians, with whom we can carry on intelligent conversations about politics and philosophy and even art and who—it never even *occurs* to us to doubt this—experience pain when they are injured and pleasure when they relax at the end of a hard day in their sulfuric-acid baths. But there is nothing inside their chitinous shells resembling a human brain: there is only purple goo bearing no resemblance whatever, even on the chemical level, to any human tissue. Now suppose physicalism is correct. If that is so, and if we really do think and feel, then our thoughts and feelings are identical with certain physical processes that go on within our brains. But obviously none of the physical processes that go on in the grey matter inside our heads goes on in the purple Scorpian goo.

Suppose, for example, that when one feels pain this event is identical with the firing of C-fibers in one's brain; pain (according to physicalism) has turned out to be the firing of C-fibers, just as bolts of lightning turned out to be massive electrical discharges and water turned out to be H_2O. But there are no C-fibers, or anything remotely resembling them, inside the Scorpians. And, therefore, if pain is the firing of C-fibers, the Scorpians do not experience pain—just as, if there is no H_2O on their planet, there is no water on their planet. It would therefore seem that if physicalism is true, neither the Scorpians nor any other beings radically unlike us in their physical structure can think and feel. Only a being that was either human or very similar to a human being could think and feel. But this conclusion can only be regarded as human (or mammalian or carbon) chauvinism. In any case, it is absurd.

A physicalist might well respond to this argument with a question: What makes you so sure it is possible for there to be creatures radically different from us in

their physical structure and capable of thought and sensation? And it might not be easy to answer this question unless bluster about chauvinism counts as an answer. But there are two replies available to the physicalist that are consistent with the assumption that the possibility of beings like the Scorpians is a real one.

Each of these replies depends upon a distinction between *types* of events and *tokens* (that is, particular instances) of those types. This distinction is best introduced by example. *War* is a type of event (or an event-type, as philosophers sometimes say), and the First World War and the Seven Years' War and the War of the Austrian Succession are three "tokens" of this one type; Lincoln's death and Caesar's death and the death of Catherine the Great are three tokens of the event-type *death*. A particular, concrete event may be—in fact, all particular, concrete events *must* be—a token of more than one type. Thus, Lincoln's death and Caesar's death are tokens not only of the type *death* but also of the type *assassination*. But fortunately not all tokens of the former are tokens of the latter: not all deaths are deaths by assassination. If every event is a token of various types, then every mental event is a token of various types, and every physical event is a token of various types.

Making use of the type-token distinction, we may distinguish two forms of physicalism (or two forms of the identity theory): type-type physicalism and token-token physicalism. Let us first examine type-type physicalism. Consider the physical event-type *a firing of C-fibers* and the mental event-type *feeling pain*. Suppose someone says that these event-types are identical, are one and the same event-type. This person's thesis could also be put this way, if we neglect some niceties about language some philosophers will not want to neglect: the phrase 'a firing of C-fibers' and the phrase 'feeling pain' are two different names for the same event-type, just as 'water' and 'the liquid that consists of H_2O molecules' are two names for the same liquid—or just as 'the Morning Star' and 'the planet Venus' are two names for the same celestial object. Type-type physicalism is a generalization of this thesis; according to type-type physicalism, *every* mental event-type is identical with some physical event-type. (But of course only an idealist would suppose that the converse holds. Idealists aside, no one would suppose that, for example, the physical event-type *volcanic eruption* was identical with some mental event-type.)

Type-type physicalism is a very strong thesis, so strong that most physicalists decline to accept it; either it is known to be false (some physicalists will say), or at least it goes far beyond the available evidence. How (the enemies of type-type physicalism ask) can we be sure even that when identical twins experience pains that feel exactly the same, there are physical events in the brains of each that are exactly alike—or even very much alike? How can we be sure there is any such pair of physical events to be found? Shouldn't it be left to the neurophysiologists to determine whether two such events exist? Should this question be settled by metaphysicians, by philosophers who have never made any neurophysiological

investigations whatever? Fortunately (most physicalists believe) there is a weaker form of physicalism available, adherence to which does not require philosophers to become armchair neurophysiologists: *token-token* physicalism.

According to token-token physicalism, each concrete mental event (such as my suddenly experiencing a sharp pain in my left arm at noon yesterday or Tim's gradual realization that Alice has been lying to him) is identical with a concrete physical event: a particular change in the physical state of someone's brain (at least in the case of human beings). But it may well be, the token-token physicalist holds, that no mental event-type is identical with any physical event-type. Perhaps, the token-token physicalist says, when Tim gradually realizes that Alice has been lying to him and his identical twin Tom gradually realizes that Alice has been lying to *him*, each of these two events is identical with a physical change in the respective brains of Tim and Tom, but these two physical changes bear little resemblance to each other (for example, it may be that they take place in different regions in the cerebral cortex). Token-token physicalism does not go so far as positively to deny that there are mental event-types that are identical with physical event-types; it simply refrains from asserting that such identities exist. If there are such identities, the token-token physicalist tells us, it is the business of observational sciences like psychology and neurophysiology to establish them; they are no more to be embraced on purely metaphysical grounds than are the chemical and astronomical identities mentioned above.

If token-token physicalism is correct, there is no problem in principle in saying, for example, that a Scorpian experiences a sensation very like the pain Jane experiences when she has a migraine. Jane's sensation of pain is, or let us suppose it is, identical with a certain pattern of C-fiber firings in her brain; the Scorpian's sensation is identical with some physical process that takes place in a reservoir of purple goo in the Scorpian's metathorax, a process that in none of its physical characteristics resembles the firing of C-fibers in a human brain.

This is the picture provided by token-token physicalism. There are many analogies that token-token physicalists have employed to make this picture a plausible one. The following analogy is typical of these. Suppose three radios are simultaneously receiving the same broadcast. One is an antique crystal set, one a vacuum-tube (valve) radio from the 1950s, and the third the latest thing in solid-state technology. We may list three "reception events": radio A's receiving the CBS broadcast of the State of the Union Message, radio B's receiving this same broadcast, and finally, radio C's receiving it. Each of these reception events is identical with a physical process going on inside one of the three radios, but the three physical processes are very different from one another. The thesis of "reception physicalism" may be defined as the thesis that reception events are physical events that go on inside radios. The thesis of type-type reception physicalism is the thesis

that each reception event-type (like *receiving the CBS broadcast of the State of the Union Message*) is identical with some physical event-type. The thesis of token-token reception physicalism is the thesis that each reception event-token, or concrete event (like *radio B's receiving the CBS broadcast of the State of the Union Message yesterday*), is identical with some concrete physical event. No doubt everyone will accept token-token reception physicalism. But the fact that the physical events that go on inside a vacuum tube are quite different from the physical events that go on inside whatever the latest solid-state devices are called renders type-type reception physicalism at best doubtful.

Doubtful, perhaps, but not wholly indefensible. I said above that there were two replies available to the physicalist consistent with the assumption that the possibility of thinking, feeling beings like the Scorpians is a real one. The first was to distinguish type-type and token-token physicalism, and to maintain that, whatever the problems faced by type-type physicalism, token-token physicalism is consistent with this possibility. The second reply is an argument for the conclusion that even type-type physicalism is consistent with the possibility of thinking, feeling beings radically different from us in anatomy and physiology—or at least that this may be so, that it is true for all we know.

We may note that event-types may be more or less abstract. The more abstract an event-type is, the weaker the conditions are that an event has to satisfy to be a token of that type, and the less abstract an event-type is, the stronger the conditions are that an event has to satisfy to be a token of that type. Here are five event-types arranged in order of decreasing abstraction: *death*, *killing* (an untimely death caused by an external agency), *murder* (a deliberate and wrongful killing of one human being by another), *assassination* (the murder of a public figure from a political motive), and *terrorist assassination* (an assassination undertaken to create a politically useful climate of fear within some group). A defender of type-type physicalism could argue that the most that the example of the Scorpians shows is that if each mental event-type is identical with some physical event-type, then the physical event-types that figure in the identities must be much more abstract than, say, *a firing of C-fibers*.

Let us return to our "radio" analogy to illustrate this idea. If we think about it, we can see that it is possible to think of a highly abstract physical event-type that has a token in each of the three radios and can plausibly be identified with the reception event-type *receiving broadcast X*. Something like this: *containing some component or components that vibrate in a way determined by the information contained in the radio waves that carry broadcast X, this vibration being amplified to the point at which it generates sound waves that are audible to the human ear*. And it seems at least somewhat plausible to suppose that something similar could be said for the case of our thoughts and feelings and those of the Scorpians. Perhaps there is some

very abstract physical event-type that is identical with, for example, the event-type *feeling pain* and which—being so very abstract—is capable of being "tokened in" both human grey matter and Scorpian purple goo. Perhaps, indeed, every mental event-type is identical with some very abstract physical event-type. Whether or not this defense of (the possibility of) type-type physicalism is correct, it seems fairly clear that physicalism cannot be refuted by an appeal to the possibility of thinking, feeling creatures radically different in their physical structure from human beings.

Suggestions for Further Reading

Chapters 2, 3, and 4, of Taylor's *Metaphysics* provide a very readable introduction to the "mind-body problem."

The two great classics of dualism are Plato's *Phaedo* and Descartes's *Meditations on First Philosophy* (see particularly Meditations II and VI).

Notes

1. The word 'materialism' is often used as a name for the thesis I am calling 'physicalism', and it has stronger and weaker senses corresponding to the stronger and weaker senses of 'physicalism'.

2. Our definition of what it is for a certain organism to be a certain person's *body* was introduced in connection with our exposition of dualistic interactionism. This definition presupposes that *x* can cause changes in *x*'s body, and that *x*'s body can cause changes in *x*. The physicalist who wants to retain the word 'body' might prefer a slightly different definition. Perhaps the physicalist would prefer to say that *x*'s body is that organism in which *x* can *bring about* changes without bringing about changes in any other organism, and it is the organism changes in which can *result in* changes in *x* without resulting in changes in any other organism. This way of wording the definition does not carry the implication that a person and that person's body are distinct things. And this way of wording the definition should be acceptable to the dualist as well, since it does not carry the implication that a person and that person's body are the same thing.

3. I thank John Keller for suggesting an improvement in an earlier definition of 'physical change'.

4. There is a position in the philosophy of mind called *property dualism*, according to which a physical thing (a human person, for example) might acquire the property "being elated" and yet this thing's acquisition of that mental property not be the same event as any physical change in the person. A discussion of property dualism is beyond the scope of this chapter.

5. It should be noted that not all theories pertaining to the relation of the human person to the human organism are either physicalistic or dualistic. We have remarked that some idealists might say there were no physical things and hence no human organisms; if there are no human organisms, there is no problem about how human organisms are related to

human persons. Some "eliminative physicalists" and "behaviorists" and some epiphenomenalists might be understood as maintaining that there are no human persons—that there is nothing for any use of the word 'I' to refer to—and thus that there is no problem about how human persons are related to human organisms. "Property dualism" (note 4) cannot easily be classified as either physicalistic or dualistic. And there are theories according to which human persons are neither physical things nor non-physical things, but are rather what we earlier called "amalgams": individual things having both physical things and non-physical things as parts. (Saint Thomas Aquinas defended a theory of this sort.) We do not have the space to discuss all these interesting theories. We shall simply assume that there are both human persons and human organisms—an assumption that leaves it an open question whether the persons and the organisms are identical. And much of our discussion of whether human persons are physical things will be relevant to the question whether every part of a human person is a physical thing.

6. G. W. Leibniz, *Monadology* (1714), §17. The translation in the text is taken from the note "Mill" in Jonathan Bennett and Peter Remnant's translation of Leibniz's *Nouveaux essais* (*New Essays on Human Understanding*, Cambridge: Cambridge University Press, 1981), lv.

7. What about *physical* simples? Could *they* think and feel? This question would not have troubled Leibniz, who thought all simples were non-physical things. (But this is a rather misleading statement if it is read without reference to the whole of his metaphysic.) The Greek atomists, however, believed that what they called atoms were physical simples, and current physics strongly suggests that various physical things—electrons, for example—have no parts. Any dualist who accepts the thesis that there are physical simples, whether in its ancient or its modern form, will probably want to say that though being without parts is a *necessary* condition of the capacity for thought and sensation, it is not sufficient; no dualist, I suppose, would be willing to say an electron was capable of thought.

8. We have defined a physical property as a property that could be possessed by, and could be possessed *only* by, a physical thing. Let us define a *non-physical* property as a property that could be possessed by, and could be possessed *only* by, a non-physical thing. (A warning about terminology: "property dualists"—see note 4—use the term 'non-physical property' in a different sense from this, and in fact in a sense incompatible with this, since according to property dualism, a physical thing *can* have "non-physical" properties.) It is important to note that just as there may be individual things that are neither physical nor non-physical, there may be properties that are neither physical nor non-physical: properties that could be possessed *either* by physical or non-physical things. (And if there are amalgams, there will be properties—such as *being an amalgam*—that can be possessed only by amalgams and are thus neither physical nor non-physical.) For example—assuming it is possible for there to be non-physical individual things—the property of being an individual thing and the property of being either physical or non-physical are both properties that are neither physical nor non-physical. Other examples would be more controversial: I think mental properties are neither physical properties nor non-physical properties, but Descartes would say they were non-physical properties, and some physicalists would say they were physical properties. (It is important to remember that a mental property is not *by definition* a non-physical property. Typical dualists believe that mental properties—properties implying thought or sensation—are non-physical properties, because they believe these

properties could be possessed by, and could be possessed only by, non-physical things. But physicalists believe that mental properties are *not* non-physical properties—because they believe that these properties could be, and in fact are, possessed by physical things.)

These considerations show that if the explanation given in the text of what Descartes meant by saying that our essence was thinking is right, then even on the assumption that we human persons are non-physical thinkers, our essence is *not* thinking. It cannot be that the only intrinsic properties a human person has or could have are mental, for *being an individual thing* is an intrinsic property of human persons, and it is not a mental property. Might we say (as I proposed in the first edition of this book) that our essence is thinking just in the case that the only *non-physical* intrinsic properties a human person has or could have are mental? This does not solve the problem, for as Alvin Plantinga has pointed out to me, *being a non-physical thing* is a non-physical property (it can be had only by non-physical things), an intrinsic property, and not a mental property (or not obviously so: perhaps *being a non-physical thing* somehow "implies either thought or sensation," but if this is the case, it is not obvious); and, according to dualism, *being a non-physical thing* is a property of human persons. At this point, I see no satisfactory explanation of the meaning of 'our essence is thinking'—I mean I see no way of explaining this phase on which, given that we are non-physical thinking things, it "comes out true" that our essence is thinking.

This completes our examination of arguments against physicalism. We now turn to arguments for physicalism. There are, I believe, four good arguments for physicalism. Like all philosophical arguments, these arguments are not decisive. To my mind, however, they tip the scale in favor of physicalism. (I do not distinguish between arguments for physicalism and arguments against dualism, since to my mind physicalism and dualism are the two most plausible theories about our nature, and an argument against dualism—unless it also tells against physicalism—is therefore an argument for physicalism.)

First, there is the *interaction argument*. We briefly mentioned in Chapter 10 some difficulties with the idea that a non-physical thing could affect a physical thing. Wouldn't that require a violation of well-established physical conservation laws like the law of the conservation of energy and the law of the conservation of linear momentum? And isn't it also far from clear how a physical thing could affect a non-physical thing? Here is another sort of "interaction" difficulty. The World, the dualist says, contains both non-physical persons and physical organisms. But how do a particular person and a particular organism become "associated"? What brings it about that Jane Tyler interacts with *this* human organism (the one we label 'Jane Tyler's body' precisely because it is the one she interacts with)? The interaction argument comprises these difficulties, together with the observation that by far the most plausible form of dualism is dualistic interactionism.

Secondly, there is the *argument from common speech*. We usually talk and act as if we were visible and tangible. We say things like, "I didn't like the way he was looking at me," or "She reached for the seat belt and buckled herself in." We don't say, "She caused her body's hands to reach for the seat belt and buckle her body in." And, while someone might say, "I didn't like the way he was looking at my body," this would mean something rather special (perhaps, 'I thought he was exhibiting undue sexual interest in me') and it couldn't always be substituted for 'I didn't like the way he was looking at me'. This suggests that our concept of a human person (or our concept of ourselves) is the concept of a thing possessing certain physical characteristics: we normally conceive of ourselves as things made of flesh and blood and bone and shaped roughly like statues of human beings.

Thirdly, there is an argument I like to call the *remote-control argument*. If dualism is true, our relation to our bodies is analogous to the relation of the operator of a remotely controlled device (such as a radio-controlled model airplane) to that device. Now consider Alfred, who is operating a model airplane by remote control.

Suppose that something—an unwary bird or a large hailstone—strikes a heavy blow to the model in midair. If the blow does significant damage to the model, we can expect that both the performance of the model and Alfred's ability to control the model will be impaired. But the blow will have no effect at all on *Alfred*, or no effect beyond his becoming aware of the blow or of some of its effects on the performance of the model and his ability to control it. But if Alfred's *body* were struck a heavy blow, and particularly if it were a blow to the head, this might have an effect on *him*, an effect that goes beyond his becoming aware of the blow and its damaging effects on his body and his ability to control his body: Alfred might well become unconscious.

This is just the sort of effect we should expect if Alfred were a certain human organism, for if the processes of consciousness are certain physical processes within the organism, a damaging blow might well cause those processes to cease, at least temporarily. But what effects should *dualism* lead us to expect from a blow to the body? I submit that if we are non-physical things, and if the processes of consciousness are non-physical processes that do not occur within the body, the most natural thing to expect is that (at the worst) we should lose control of our bodies while continuing to be conscious. The blow to the base of Alfred's skull that in fact produces unconsciousness should, according to dualism, produce the following effects on Alfred: he experiences a sharp pain at the base of his skull; he then notes that his body is falling to the floor and that it no longer responds to his will; his visual sensations and the pain at the base of his skull and all the other sensations he has been experiencing fade away; and he is left, as it were, floating in darkness, isolated, but fully conscious and able to contemplate his isolated situation and to speculate about its probable causes and its duration. But this is not what happens when one receives a blow at the base of the skull. One never finds oneself conscious but isolated from one's body.

Dualism, therefore, seems, on the face of it, to make wrong predictions about what the human person will experience in certain situations. Here is another wrong prediction that dualism seems to make: if dualism were correct, we should expect that the ingestion of large quantities of alcohol would result in a partial or complete loss of motor control but leave the mind clear. Physicalism, however, would predict the former effect and would also strongly suggest that the drinker's mental processes would be impaired. Because dualism makes (or seems on the face of it to make) these wrong predictions, it is doubtful. I say 'doubtful' rather than 'false', because the defender of dualism will not have too much difficulty in contriving a hypothesis to explain away the fact that a blow to the base of the skull causes one to lose consciousness or the fact that the ingestion of alcohol impairs one's mental processes. For example, the dualist might suggest that a temporary interruption of the normal causal interaction between the person and the body has

a traumatic effect on the person, a salient feature of which is loss of consciousness. But this does not change the fact that the typical effects of a blow to the base of the skull are something that has to be *explained away* by dualists and are therefore an embarrassment to them. I say 'is doubtful' rather than 'faces a difficulty' because it is my hope that the reader will find all the hypotheses by which the dualist explains away the observed effects of a blow to the base of the skull (or the ingestion of alcohol) to be implausible and ad hoc. I find them so; if I am wrong about the typical reaction of the disinterested reader to these hypotheses, I have claimed too much by using the word 'doubtful'.

Finally, there is the *duplication argument*. This is the single argument for physicalism that I find the most powerful and persuasive. Recall the "duplicating machine" we imagined in Chapter 2, in connection with our discussion of the concept of an intrinsic property. Let us imagine this machine and its operations in a little more detail. The duplicating machine consists of two chambers connected by an impressive mass of science-fictional gadgetry. If you place any physical object inside one of the chambers and press the big red button, a perfect physical duplicate of the object appears in the other chamber. The notion of a perfect physical duplicate may be explained as follows. A physical thing is composed entirely of quarks and electrons. A perfect physical duplicate of the physical thing *x* is a thing composed entirely of quarks and electrons arranged in the same way in relation to one another as the quarks and electrons composing *x* are, and each of the quarks and electrons composing a perfect physical duplicate of *x* will be in the same physical state as the corresponding particle in *x*. If, for example, you place the Koh-i-Noor diamond in one of the chambers and press the button, a thing *absolutely indistinguishable from* the Koh-i-Noor (since it is a perfect physical duplicate of the Koh-i-Noor) will appear in the other. If the two objects are placed side by side and then moved in a rapid and confusing way, so that everyone loses track of which was the original and which the duplicate, no one, no jeweler, mineralogist, or physicist, will ever be able to tell, by any test whatever, which of the two played an important role in the history of the British Raj in the nineteenth century and which was created a moment ago in the duplicating machine.

Now let us consider a second case of duplication. A marble is slowly rolling across the floor of one of the chambers. The button is pressed. There appears on the floor of the other chamber a marble of the same shape and size and weight and color, rolling in the same direction and at the same speed: our machine reproduces not only the "static" properties of a thing, but also its "dynamic" properties.

Now let us place a living mouse in the chamber and press the button. What will appear in the other chamber? Another living mouse, surely? And wouldn't it be a mouse in every respect interchangeable with the original? If, for example, the original mouse had been taught to get cheese from a cheese dispenser by pressing a lever

when a light flashed, wouldn't the new mouse know this trick, too? Knowledge of how and when to press the lever to get cheese must somehow be stored in the mouse's little brain, and since the duplicate mouse's brain is a *perfect* duplicate of the original's brain, right down to the subatomic level, the same knowledge must be stored in the duplicate brain. (If you used the machine to duplicate a flash drive in which you'd stored the novel you'd written, you wouldn't get a "blank" flash drive in the other chamber; you'd get another flash drive that contained the novel: in duplicating *every* physical characteristic of the original, the machine automatically duplicates those characteristics of the flash drive that encode a record of the sequence of keystrokes that form your novel.)

And now, finally, let us put *Alfred* into one of the chambers of the duplicating machine and press the button. What do we find in the other chamber? A very intelligent Muslim student of mine once assured me that what one would find would be a dead human body—since the duplicating machine would not reproduce Alfred's soul, which was the principle of life. This dead body, at the instant of its appearance, would be standing just as Alfred stood, and on its face would be an expression just like the expression on Alfred's face. Even in that first instant, however, the body would not be alive, and having appeared, it would immediately collapse and lie unmoving, its face the blank mask of a corpse. (As a testimony to the general intellectual capacity of my student, I will mention that he was the salutatorian of his graduating class and went on to earn a Ph.D. in nuclear engineering.) I think Plato would have agreed with my student. Descartes, however, would not have agreed. Descartes would have contended that a *living* human body would have appeared in the other chamber. But, Descartes would have said, this body would immediately crumple to the floor. It would then lie there breathing and perhaps drooling, and, if you force-fed it, it would digest the food and in time produce excreta. But it would not *do* anything much. It would just lie there breathing and drooling and digesting and excreting. And this, of course, would be because there was no mind or soul or person in interaction with it. As a consequence, no thought or sensation would be in any way associated with the duplicate body. Life, in the strict, biological sense, was for Descartes (as it was not for Plato or for my student) a purely physical phenomenon; thought and sensation were not. Modern molecular biology, I think, has shown that Descartes was right about life—or has at least rendered the thesis that life is a complex physical process vastly more probable than its denial. But what about thought and sensation?

That is the question. It is essentially the question whether physicalism is true. The story of the duplicating machine is a device to focus our thoughts as we consider this question. Dualists must say that since thought and sensation are not physical processes occurring within a living human organism, the human body the duplicating machine creates will crumple mindlessly, just as Descartes would

have predicted. (I doubt whether many people raised and educated in a European or "European-descended" culture would agree with my Muslim student that the duplicating machine would produce a corpse.) But is this really what any of us believes? Aren't we strongly inclined to believe—at least when we are not considering the consequences of what we believe for the metaphysics of the human person—that the duplicate would "have" thoughts and feelings and beliefs and memories (or what felt like memories; they would not, of course, be connected with past events in the way a real memory is) and desires and emotions? Aren't we strongly inclined to believe that the duplicate would have a conscious mental life like our own and would display the content of this conscious mental life in his observable behavior?

Those who do believe this will concede, after a moment's reflection, that just as most of the duplicate's memories will not be real memories, so most of his beliefs about himself and his history will be false. The duplicate will, for example, believe that he is Alfred, and he is not. That is, he is not a man who has existed for such-and-such a number of years (he is only a few minutes old) and is married to Winifred (he has never met her), and so on. The duplicate is in no sense Alfred. He is someone else, for if you stick a pin into Alfred, the duplicate feels no pain. Nevertheless, it *seems* to the duplicate that he is Alfred. What it is *like* to be the duplicate is just exactly what it is like to be Alfred. If Alfred was unconscious when he was duplicated, and if he and the duplicate were then "scrambled" (like the two diamonds in our earlier example), no one, including Alfred and the duplicate, could ever know which was Alfred and which was the duplicate. Alfred himself would have to say—at least if he were fully, and perhaps inhumanly, reasonable—"For all I know, I am the duplicate." And if by some chance it were the duplicate that went home to Winifred, she would never suspect that he was not her husband. And just as Winifred would never suspect that anything was amiss, neither would Alfred's children or his mother or his closest friend or his confessor or his psychiatrist.

If this were indeed the outcome of running Alfred through the duplicating machine, dualism would be effectively refuted. The dualist *could*—this sort of thing is almost always possible—contrive some hypothesis that would explain away this outcome. The dualist might, for example, propose that whenever a human body is perfectly duplicated, God creates a perfect duplicate of the non-physical person who had been interacting with the original body and so arranges matters that the duplicate person is in interaction with the duplicate body. But this would be a desperate move. It would be far more reasonable, even for theists, to conclude that the observed result of our "experiment" should be explained as follows: the thoughts and feelings of a human person are physical processes within a human organism, and in making a perfect physical duplicate of a human organism, we produce a

human organism with the same thoughts and feelings. (The same, that is, at the first moment of the new organism's existence. The thoughts and feelings of the two organisms would probably diverge almost immediately, since they would probably find themselves almost immediately in different situations.) It would be reasonable to conclude that the mental properties of a human person are related to the physical properties of that person in a way analogous to the way in which the software associated with a particular computer is related to the physical properties of that computer.

The fact that certain software is associated with (is present in, has been programmed into, is embodied by) a particular computer is as much a physical fact about that computer as are any facts about the hardware constituting the "architecture" of that computer. If I were to take the laptop with which I am writing these words and place it—while it is up and running—in the duplicating machine, the computer the machine produced would not be simply another computer of (apparently) the same make and model; immediately after the duplication, the same words would be visible on its screen, and, like the original, it would (apparently) be running Microsoft® Word for Mac 2011, 14.3.9, and it would respond in exactly the same way as the original to anything done at the keyboard.

And we have—don't we?—a strong tendency to believe that duplicating a living human organism would have the analogous result as regards the mental life of the human person whose body that organism is: just as, in making a perfect physical duplicate of a working computer, we duplicate all the software programmed into that computer, so, in making a perfect physical duplicate of a living human organism, we duplicate the entire psychology associated with that organism—everything from a neurotic fear of snakes and the ability to speak Russian to a hardly noticeable pain in the left elbow.

Anyone who can honestly reply to this question by saying something along the lines of, "Well, *I* don't observe any such tendency in myself. Like Descartes, I think the duplicate would crumple and fall to the floor and drool," will not be moved by the duplication argument. Anyone who, on reflection, decides that the duplicate would exhibit behavior indistinguishable from Alfred's (in the same situations) should conclude that the duplicate has a mental life like Alfred's and that physicalism is therefore true and dualism false.

This concludes our discussion of the nature of rational beings—or at any rate, of human beings, the only rational beings whose existence is uncontroversial. This discussion has been highly tentative. We should remember that even if we have succeeded in showing that physicalism is the most reasonable theory about the nature of human beings, we have not done anything to dispel the mystery of that nature. Thought and feeling remain as we found them: impenetrable mysteries.

Suggestions for Further Reading

There are two excellent collections of essays devoted to the problem of personal identity: Perry's *Personal Identity* and Rorty's *The Identities of Persons*. For Judith Jarvis Thomson's reasons for thinking that an explanation of identity across time in terms of four-dimensional objects constitutes "a crazy metaphysic," see her "Parthood and Identity across Time" (a very difficult essay for those who are not formally trained in philosophy). The idea that there is a close analogy between computer hardware and software, on the one hand, and the physical and mental aspects of human beings, on the other, has been extremely influential in philosophy since about the middle of the 1960s. Parts II, III, and IV of Hoffstadter and Dennett's *The Mind's I* provide an excellent introduction to the use philosophers have made of this fascinating idea.

Notes

1. The three terms 'the Original Ship', 'the Reconstructed Ship', and 'the Continuous Ship' were invented by Jonathan Bennett.

2. This argument is not watertight, even given that a physical thing cannot survive a change of parts. A physicalist *could* maintain that we are physical simples, or that each of us is some composite but very small thing (presumably located inside the brain) that does not change its parts. In fact, one physicalist has maintained this. But few have found it an attractive position. If *I* were convinced that the only way to render physicalism consistent with personal identity across time was to postulate that each person was a tiny object inside that person's brain, I'd become a dualist. (As, eventually, did the physicalist to whom I have alluded.)

3. J. Z. Young, *An Introduction to the Study of Man* (Oxford: The Clarendon Press, 1971), 86–87.

4. In the Hebrew Bible (Daniel 12:2) we read "And many of them that sleep in the dust of the earth shall awake, some to everlasting life and some to shame and everlasting contempt." The Christian "Athanasian Creed" speaks of the resurrection of the dead in these words: "All human beings shall rise again with their bodies and shall give account for their own works. . . ."