entailment, it is not at all easy to accept the entailment in the case of thinking that it is raining. It is one of Sellars' remaining tasks to convince us that we should accept it.

JAMES W. CORNMAN

THE UNIVERSITY OF ROCHESTER

SYMPOSIUM: MINDS AND MACHINES

ROBOTS: MACHINES OR ARTIFICIALLY CREATED LIFE?*

THOSE of us who passed many (well- or ill-spent?) childhood hours reading tales of rockets and robots, androids and telepaths, galactic civilizations and time machines, know all too well that robots—hypothetical machines that simulate human behavior, often with an at least roughly human appearance—can be friendly or fearsome, man's best friend or worst enemy. When friendly, robots can be inspiring or pathetic—they can overawe us with their superhuman powers (and with their greater than human virtue as well, at least in the writings of some authors), or they can amuse us with their stupidities and naivete. Robots have been "known" to fall in love, go mad (power- or otherwise), annoy with oversolicitousness. At least in the literature of science fiction, then, it is possible for a robot to be "conscious"; that means (since 'consciousness', like 'material object' and 'universal', is a philosopher's stand-in for more substantial words) to have feelings, thoughts, attitudes, and character traits. But is it really possible? If it is possible, what are the necessary and sufficient conditions? And why should we philosophers worry about this anyway? Aren't the mind-body problem, the problem of other minds, the problem of logical behaviorism, the problem: What did Wittgenstein really mean in the private-language argument? (and why should one care?), more than enough to keep the most industrious philosopher of mind busy without dragging in or inventing the Problem of the Minds of Machines?—These are my concerns in this paper.

The mind-body problem has been much discussed in the past thirty-odd years, but the discussion seems to me to have been fruitless. No one has really been persuaded by *The Concept of Mind* that the relation of university to buildings, professors, and stu-

* To be presented in a symposium on "Minds and Machines" at the sixty-first annual meeting of the American Philosophical Association, Eastern Division, December 28, 1964.

dents is a helpful model for the relation of mind to body, or even for the relation of, say, *being intelligent* to individual speech-acts. And Herbert Feigl informs me that he has now himself abandoned his well-known "identity theory" of the mind-body relation. The problem of other minds has been much more fruitful—the well-known and extremely important paper by Austin is ample testimony to that—but even that problem has begun to seem somewhat stale of late. What I hope to persuade you is that the problem of the Minds of Machines will prove, at least for a while, to afford an exciting new way to approach quite traditional issues in the philosophy of mind. Whether, and under what conditions, a robot could be conscious is a question that cannot be discussed without at once impinging on the topics that have been treated under the headings Mind-Body Problem and Problem of Other Minds. For my own part, I believe that certain crucial issues come to the fore almost of their own accord in this connection—issues which *should* have been discussed by writers who have dealt with the two headings just mentioned, but which have not been—and, therefore, that the problem of the robot becomes almost obligatory for a philosopher of mind to discuss.

Before starting I wish to emphasize, lest any should misunderstand, that my concern is with how we should speak about humans and not with how we should speak about machines. My interest in the latter question derives from my just-mentioned conviction: that clarity with respect to the "borderline case" of robots, if it can only be achieved, will carry with it clarity with respect to the "central area" of talk about feelings, thoughts, consciousness, life, etc.

Minds and Machines

In an earlier paper,[1] I attempted to show that a problem *very* analogous to the mind-body problem would automatically arise for robots. The same point could easily have been made in connection with the problem of other minds. To briefly review the argument: conceive of a community of robots. Let these robots "know" nothing concerning their own physical make-up or how they came into existence (perhaps they would arrive at a robot Creation Story and a polytheistic religion, with robot gods on a robot Olympus). Let them "speak" a language (say, English), in conformity with the grammatical rules and the publicly observable semantic and discourse-analytical regularities of that language.

[1] "Minds and Machines," in Sidney Hook, ed., *Dimensions of Mind* (New York: NYU Press, 1960), pp. 148–179.

What might the role of psychological predicates be in such a community?

In the paper referred to, I employed a simple "evincing" model for such predicates. Since this model is obviously *over*-simple, let us tell a more complicated story. When a robot sees something red (something that evokes the appropriate internal state in the robot) he calls it "red." Our robots are supposed to be capable of inductive reasoning and theory construction. So a robot may discover that something he called red was not really red. Then he will say "well, it looked red." Or, if he is in the appropriate internal state for red, but knows on the basis of cross-inductions from certain other cases that what he "sees" is not really red, he will say "it *looks* red, but it isn't really red." Thus he will have a distinction between the physical reality and the visual appearance, just as we do. But the robot will never say "that looks as if it looked red, but it doesn't really look red." That is, there is no notion in the robot-English of an *appearance of an appearance of red,* any more than there is in English. Moreover, the reason is the same: that any state which cannot be discriminated from "looks-red" *counts* as "looks-red" (under normal conditions of linguistic proficiency, absence of confusion, etc.). What this illustrates, of course, is that the "incorrigibility" of statements of the form "that looks red" is to be explained by an elucidation of the logical features of such discourse, and not by the metaphor of "direct" access.

If we assume that these robots are unsophisticated scientifically, there is no reason for them to know more of their own internal constitution than an ancient Greek knew about the functioning of the central nervous system. We may imagine them developing a sophisticated science in the course of centuries, and thus eventually arriving at tentative identifications of the form: "when a thing 'looks red' to one of us, it means he is in internal state 'flip-flop 72 is on'." If these robots also publish papers on philosophy (and why should a robot not be able to do considerably better than many of our students?), a lively discussion may ensue concerning the philosophical implications of such discoveries. Some robots may argue, "*obviously,* what we have discovered is that 'seeing red' *is* being in internal state 'flip-flop 72 on' "; others may argue, "*obviously,* what you made was an *empirical* discovery; the *meaning* of 'it looks red' isn't the same as the *meaning* of 'flip-flop 72 is on'; hence the *attributes* (or states, or conditions, or properties) 'being in the state of seeming to see something red' and 'having flip-flop 72 on' are *two* attributes (or states, or conditions, or properties) and not *one*"; others may

argue "when I have the illusion that something red is present, nothing red is physically there. Yet, in a sense, I *see* something red. What I see, I *call* a sense-datum. The sense datum is red. The flip-flop isn't red. So, *obviously,* the sense-datum can't be identical with the flip-flop, on or off." And so on. In short, robots can be just as bad at philosophy as people. Or (more politely), the *logical* aspects of the Mind-Body Problem are aspects of a problem that *must* arise for any computing system satisfying the conditions that (1) it uses language and constructs theories; (2) it does not initially "know" its own physical make-up, except superficially; (3) it is equipped with sense organs, and able to perform experiments; (4) it comes to know its own make-up through empirical investigation and theory construction.

Some Objections Considered

The argument just reviewed seems extremely simple. Yet some astonishing misunderstandings have arisen. The one that most surprised me was expressed thus: "As far as I can see, all you show is that a robot could simulate human *behavior.*" This objection, needless(hopefully)-to-say, misses the point of the foregoing *completely.* The point is this: that a robot or a computing machine can, *in a sense,* follow rules (Whether it is the same sense as the sense in which a man follows rules, or only analogous, depends on whether the particular robot can be said to be "conscious," etc., and thus on the central question of this paper.); that the meaning of an utterance is a function of the rules that govern its construction and use; that the rules governing the *robot* utterances 'I see something that looks red' and 'flip-flop 72 is on' are quite different. The former utterance may be correctly uttered by any robot which has "learned" to discriminate red things from non-red things correctly, judged by the consensus of the other robots, and which finds itself in the state that signals the presence of a red object. Thus, in the case of a normally constructed robot, 'I see something that looks red' may be uttered whenever flip-flop 72 is on, *whether the robot "knows" that flip-flop 72 is on or not.* 'Flip-flop 72 is on' may be correctly (reasonably) uttered only when the robot "knows" that flip-flop 72 is on —i.e., only when it can *conclude* that flip-flop 72 is on from empirically established theory together with such observation statements as its conditioning may prompt it to utter, or as it may hear other robots utter. 'It looks red' is an utterance for which it does not and cannot give reasons. 'Flip-flop 72 is on' is an utterance for which it can give reasons. And so on. Since these

semantic differences are the same for the robot as for a human, any argument from the semantic nonequivalence of internal(physical)-state statements and "looks" statements to the character of mind or consciousness must be valid for the robot if it is valid for a human. (Likewise the argument from the alleged fact that there is "a sense of *see*" in which one can correctly say "I see something red" in certain cases in which nothing red is physically present.)

Besides the misunderstandings and nonunderstandings just alluded to, some interesting objections have been advanced. These objections attempt to break the logical analogy just drawn by me. I shall here briefly discuss two such objections, advanced by Prof. Kurt Baier.

Baier's first argument[2] runs as follows: The connection between my visual sensation of red and my utterance 'it looks as if there is something red in front of me' (or whatever) is *not* merely a causal one. The sensation does not *merely* evoke the utterance; I utter the utterance because I *know* that I am having the sensation. But the robot utters the utterance because he is *caused* to utter it by his internal state (flip-flop 72 being on). Thus there is a fundamental disanalogy between the two cases.

Baier's second argument is as follows: Certain *qualia* are *intrinsically* painful and others are *intrinsically* pleasurable. I cannot conceive of an intrinsically unpleasant quale Q being exactly the same for someone else "only he finds it pleasurable." However, if a robot is programmed so that it *acts as if* it were having a pleasant experience when, say, a certain part of its anatomy jangles, it could easily be reprogrammed so that it would act as if it were having a painful, and not a pleasant, experience upon those occasions. Thus the counterparts of "qualia" in the robot case—certain physical states—lack an essential property of qualia: they cannot be *intrinsically* pleasurable or painful.

Can a robot have a sensation? Well, it can have a "sensation." That is, it can be a "model" for any psychological theory that is true of human beings. If it is a "model" for such a theory, then when it is in the internal state that corresponds to or "realizes" the psychological predicate "has the visual sensation of red," it will act as a human would act (depending also on what other "psychological" predicates apply). That is, "flip-flop 72 being on" does not have to *directly* (uncontrollably) "evoke" the utterance 'It looks as if there is something red in front of me'. I agree with Baier that so simple an "evincing" model will certainly

[2] These arguments come from an unpublished paper by Baier, which was read at a colloquium at the Albert Einstein College of Medicine in 1962.

not do justice to the character of such reports—but not in the case of robots either!

What is it for a person to "know" that he has a sensation? Since only philosophers talk in this way, no uniform answer is to be expected. Some philosophers identify having a sensation and knowing that one has it. Then "I know I have the visual sensation of red" just means "I have the visual sensation of red," and the question "Can the robot *know* that he has the 'sensation' of red?" means "Can the robot have the 'sensation' of red?"—a question which we have answered in the affirmative. (I have not argued that "sensations" are *sensations,* but only that a thoroughgoing logical analogy holds between sensation-talk in the case of humans and "sensation"-talk in the case of robots.) Other philosophers (most recently Ayer, in *The Concept of a Person*) have argued that to *know* one has a sensation one must be able to describe it. But in this sense, too, a robot can know that he has a "sensation." If knowing that p is having a "multi-tracked disposition" to appropriate sayings and question-answerings and behavings, as urged by Ryle in *The Concept of Mind,* then a robot can know anything a person can. A robot, just as well as a human, could participate in the following dialogue:

A. Describe the visual sensation you just mentioned.
B. It is the sensation of a large red expanse.
A. Is the red uniform—the same shade all over?
B. I think so.
A. Attend carefully!
B. I am!

Unfortunately for this last argument, Ryle's account of knowing is incorrect; no specifiable disposition to sayings and behavings, "multi-tracked" or otherwise, can *constitute* a knowing-that in the way in which certain specifiable arrangements and interrelationships of buildings, administrators, professors, and students will constitute a university. "Knowing that," like being in pain and like preferring, is only mediately related to behavior: knowing-that p involves being disposed to answer certain question correctly *if I want to, if I am not confused,* etc. And wanting to answer a question correctly is being disposed to answer it correctly *if I know the answer, if there is nothing I want more,* etc.—Psychological states are characterizable only in terms of their relations to each other (as well as to behavior, etc.), and not as dispositions which can be "unpacked" without coming back to the very psychological predicates that are in question. But this is not fatal to our case: A robot, too, can have internal states that are related to each other (and only indirectly to behavior and sensory stimula-

tion) as required by a psychological theory. Then, when the robot is in the internal state that realizes the predicate "knows that p" we may say that the robot "knows" that p. Its "knowing" may not be *knowing*—because it may not "really be conscious"—that is what we have to decide; but it will play the role in the robot's behavior that *knowing* plays in human behavior. In sum, for any sense in which a human can "know that he has a sensation" there will be a logically and semantically analogous sense in which a robot can "know" that he has a "sensation." And this is all that my argument requires.

After this digression on the logical character of "knowing," we are finally ready to deal with Baier's first argument. The argument may easily be seen to be a mere variant of the "water-on-the-brain" argument (you can have water on the brain but not water on the mind; hence the mind is not the brain). One can know that one has a sensation without knowing that one is in brain-state S; hence the sensation cannot be identical with brain-state S. This is all the argument comes to. But, since "knowing that" is an intensional context, a robot can correctly say "I don't know that flip-flop 72 is on (or even what a 'flip-flop' is, for that matter)," even in situations in which it can correctly assert, "I have the 'sensation' of red." It can even assert: "I 'know' that I have the 'sensation' of red." If it follows in the human case that the sensation of red is not identical with the brain-state S, then by the same argument from the same semantical premises, the robot philosophers can conclude that the "sensation" of red is not identical with "flip-flop 72 being on." The robot philosopher too can argue: "I am not merely *caused* to utter the utterance 'It looks as if there is something red in front of me' by the occurrence of the 'sensation'; part of the causation is also that I '*understand*' the words that I utter; I 'know' that I am having the 'sensation'; I 'wish' to report my 'sensation' to other robots; etc." And, indeed, I think that Baier and the robot philosopher are both right. Psychological attributes, whether in human language or in robot language, are simply *not* the same as physical attributes. To say that a robot is angry (or "angry") is a quite different predication from the predication "such and such a fluid has reached a high concentration," even if the latter predicate "physically realizes" the former. Psychological theories say that an organism has certain states which are *not* specified in "physical" terms, but which are taken as primitive. Relations are specified between these states, and between the totality of the states and sensory inputs ("stimuli") and behavior ("responses").

Thus, as Jerry Fodor has remarked,[3] it is part of the "logic" of psychological theories that (physically) *different* structures may obey (or be "models" of) the *same* psychological theory. A robot and a human being may exhibit "repression" or "inhibitory potential" in exactly the same sense. I do not contend that 'angry' is a primitive term in a psychological theory; indeed, this account, which has been taken by some as a reaction to Ryle-ism, seems to me to create puzzles where none should exist (if 'angry' is a theoretical term, then "I am angry" must be a *hypothesis!*); but I do contend that the patterns of correct usage, in the case of an ordinary-language psychological term, no more presuppose or imply that there is an *independently* specifiable state which "realizes" the predicate, or, if there is one, that it is a *physical* state in the narrow sense (definable in terms of the vocabulary of present-day physics), or, if there is one, that it is the *same* for all members of the speech community, than the postulates of a psychological theory do. Indeed, there could be a community of robots that did *not* all have the same physical constitution, but did all have the same *psychology;* and such robots could *univocally* say "I have the sensation of red," "you have the sensation of red," "he has the sensation of red," even if the three robots referred to did not "physically realize" the "sensation of red" in the same way. Thus the *attributes:* having the "sensation" of red and "flip-flop 72 being on" are simply *not* identical in the case of the robots. If Materialism is taken to be the denial of the existence of "nonphysical" attributes, then Materialism is false even for robots!

Still, Baier might reply: if I say that a robot has the "sensation" of red, I mean that he is in *some* physical state (a "visual" one) that signals to him the presence of red objects; if I say that a human has the sensation of red, I do not mean that he is necessarily in some special *physical* state. *Of course,* there is a *state* I am in when and only when I have the sensation of red—namely, the state of having a sensation of red. But this is a remark about the logic of 'state', and says *nothing* about the meaning of 'sensation of red'.

I think that this is right. When *we* say: "that robot has the 'sensation' of red," there are (or would be) implications that are not present when we talk about each other. But that is because we think of the robots *as* robots. Let us suppose that the robots do *not* "think" of themselves as robots; according to their theory,

[3] "Psychological Explanation," to appear in a forthcoming collection edited by Max Black.

they have (or possibly have) "souls." Then, when a robot says of another robot "he has the 'sensation' of red" (or something in more ordinary language to this effect), the implication will *not* be present that the other robot must be in any special *physical* state. Why should it not be an open possibility for the robot scientists and philosophers that they will *fail* to find "correlates" at the physical level for the various sensations they report, just as it is an open possibility for us that we will fail to find such correlates? To carry the analogy one final step further: if the robots go on to manufacture ROBOTS (i.e., robots that the robots themselves regard as *mere* robots), a robot philosopher will sooner or later argue: "when I say that a ROBOT 'thinks that something is red', or that something 'looks red' to a ROBOT, all that I mean is that the ROBOT is in a certain kind of *physical* state (admittedly, one specified by its *psychological* significance, and not by a direct physical-chemical description). The ROBOT must be able to discriminate red from non-red things, and the state in question must figure in a certain rather-hard-to-describe way in the discrimination process. But when I say that a fellow *person* (robot) 'thinks that something is red,' etc., I do not mean that he is necessarily in any special kind of physical state. Thus, in the only philosophically interesting sense of 'sensation,' persons (robots) have 'sensations' and ROBOTS do not." I conclude that Baier's first argument does not break my analogy.

The second argument seems to me to rest on two dubious premises. Granted, if the physical correlate of a given painful quale Q is something peripheral, then my brain could be "reprogrammed" so that the event would become the physical correlate of some pleasurable psychological state; if the correlate is a highly structured state of the whole brain, then such reprogramming may well be impossible. Thus the premise: Let S be the state of the robot's brain that "realizes" some "pleasure quale"; then, in principle, the robot's brain could always be reprogrammed so that S would "realize" a "painful quale" instead—seems to be simply false. (The other dubious premise is the existence of *intrinsically* pleasant and painful qualia. This is supposed to be introspectively evident, but I do not find it so.)

Should Robots Have Civil Rights?

Throughout this paper I have stressed the possibility that a robot and a human may have the same "psychology"—that is, they may obey the same psychological laws. To say that two organisms (or systems) obey the same psychological laws is not

at all the same thing as to say that their behavior is similar. Indeed, two people may obey the same psychological laws and exhibit *different* behavior, even given similar environments in childhood, partly because psychological laws are only statistical and partly because crucial parameters may have different values. To know the psychological laws obeyed by a species, one must know how *any* member of that species *could* behave, given the widest variation in all the parameters that are capable of variation at all. In general, such laws, like all scientific laws, will involve abstractions—terms more or less remote from direct behavioral observation. Examples of such terms have already been given: repression, inhibitory potential, preference, sensation, belief. Thus, to say that a man and a robot have the same "psychology" (are *psychologically isomorphic,* as I will also say) is to say that the behavior of the two *species* is most simply and revealingly analyzed, at the psychological level (in abstraction from the details of the internal physical structure), in terms of the *same* "psychological states" and the same hypothetical parameters. For example, if a human being is a "probabilistic automaton," then any robot with the same "machine table" will be psychologically isomorphic to a human being. If the human brain is simply a neural net with a certain program, as in the theory of Pitts and McCulloch, then a robot whose "brain" was a similar net, only constructed of flip-flops rather than of neurons, would have exactly the same psychology as a human. To avoid question-begging, I will consider psychology as a science that describes the behavior of any species of systems whose behavior is amenable to behavioral analysis, and interpretation in terms of molar behavioral "constructs" of the familiar kind (stimulus, response, drive, saturation, etc.). Thus, saying that a robot (or an octopus) has a *psychology* (obeys psychological laws) does not imply that it is necessarily conscious. For example, the mechanical "mice" constructed by Shannon have a psychology (indeed, they were constructed precisely to serve as a model for a certain psychological theory of conditioning), but no one would contend that they are alive or conscious. In the case of Turing Machines, finite automata, etc., what I here call "psychological isomorphism" is what I referred to in previous papers as "sameness of functional organization."

In the rest of this paper, I will imagine that we are confronted with a community of robots which (who?) are psychologically isomorphic to human beings in the sense just explained. I will also assume that "psychophysical parallelism" holds good for human beings and that, if an action can be explained psychologi-

cally, the corresponding "trajectory" of the living human body that executes that action can be explained (in principle) in physical-chemical terms. The possibility of constructing a robot psychologically isomorphic to a human being does not depend on this assumption; a robot could be psychologically isomorphic to a disembodied spirit or to a "ghost in a machine" just as well, if such there were; but the conceptual situation will be a little less confusing if we neglect *those* issues in the present paper.

Let Oscar be one of these robots, and let us imagine that Oscar is having the "sensation" of red. Is Oscar having the sensation of red? In more ordinary language: is Oscar *seeing* anything? Is he thinking, feeling anything? Is Oscar Alive? Is Oscar Conscious?

I have referred to this problem as the problem of the "civil rights of robots" because that is what it may become, and much faster than any of us now expect. Given the ever-accelerating rate of both technological and social change, it is entirely possible that robots will one day exist, and argue "we *are* alive; we *are* conscious!" In that event, what are today only philosophical prejudices of a traditional anthropocentric and mentalistic kind would all too likely develop into conservative political attitudes. But fortunately, we today have the advantage of being able to discuss this problem disinterestedly, and a little more chance, therefore, of arriving at the correct answer.

I think that the most interesting case is the case in which (1) "psychophysical parallelism" holds (so that it can at least be contended that *we* are just as much "physical-chemical systems" as robots are), and (2) the robots in question are psychologically isomorphic to us. This is surely the most favorable case for the philosopher who wishes to argue that robots of "a sufficient degree of complexity" would (not just *could,* but necessarily *would*) be conscious. Such a philosopher would presumably contend that Oscar had sensations, thoughts, feelings, etc., in just the sense in which we do and that the use of "raised-eyebrow" quotes throughout this paper whenever a psychological predicate was being applied to a robot was unnecessary. It is this contention that I wish to explore, not with the usual polemical desire to show either that materialism is correct and, hence (?), that such robots as Oscar would be conscious or to show that all such questions have been resolved once and for all by *Philosophical Investigations,* God but give us the eyes to see it, but rather with my own perverse interest in the logical structure of the quaint and curious bits of discourse that philosophers propound as "arguments"—

and with a perhaps ultimately more serious interest in the relevant semantical aspects of our language.

Anti-civil-libertarian Arguments

Some of the arguments designed to show that Oscar *could not* be conscious may be easily exposed as bad arguments. Thus, the *phonograph-record argument:* a robot only "plays" behavior in the sense in which a phonograph record plays music. When we laugh at the joke of a robot, we are really appreciating the wit of the human programmer, and not the wit of the robot. The *reprogramming argument:* a robot has no real character of its own. It could at any time be reprogrammed to behave in the reverse of the way it has previously behaved. But a human being who was "reprogrammed" (say, by a brain operation performed by a race with a tremendously advanced science), so as to have a new and completely predetermined set of responses, would no longer be a human being (in the full sense), but a monster. The *question-begging argument:* the so-called "psychological" states of a robot are in reality just physical states. But *our* psychological states are *not* physical states. So it could only be in the most Pickwickian of senses that a robot was "conscious."

The first argument ignores the possibility of robots that *learn.* A robot whose "brain" was merely a library of predetermined behavior routines, each imagined in full detail by the programmer, would indeed be uninteresting. But such a robot would be incapable of learning anything that the programmer did not know, and would thus fail to be psychologically isomorphic to the programmer, or to any human. On the other hand, if the programmer constructs a robot so that it will be a model of certain psychological laws, he will *not,* in general, know how it will behave in real-life situations, just as a psychologist might know all of the *laws* of human psychology, but still be no better (or little better) than any one else at predicting how humans will behave in real-life situations. Imagine that the robot at "birth" is as helpless as a newborn babe, and that it acquires our culture by being brought up with humans. When it reaches the stage of inventing a joke, and we laugh, it is simply not true that we are "appreciating the wit of the programmer." What the programmer invented was not a joke, but a system which could one day produce new jokes. The second argument, like the first, assumes that "programmed" behavior must be wholly predictable and lack all spontaneity. If I "reprogram" a criminal (via a brain operation) to become a good citizen, but without destroying his capacity to learn, to

develop, to change (perhaps even to change back into a criminal some day), then I have certainly not created a "monster." If Oscar is psychologically isomorphic to a human, then Oscar can be "reprogrammed" to the extent, and only to the extent, that a human can. The third argument assumes outright that psychological predicates never apply to Oscar and to a human in the same sense, which is just the point at issue.

All these arguments suffer from one unnoticed and absolutely crippling defect. They rely on just two facts about robots: that they are artifacts and that they are deterministic systems of a physical kind, whose behavior (including the "intelligent" aspects) has been preselected and designed by the artificer. But it is purely contingent that these two properties are *not* properties of human beings. Thus, if we should one day discover that *we* are artifacts and that our every utterance was anticipated by our superintelligent creators (with a small "c"), it would follow, if these arguments were sound, that *we* are not conscious! At the same time, as just noted, these two properties are *not* properties of *all* imaginable robots. Thus these arguments fail in two directions: they might "show" that *people* are *not* conscious—because people might be the wrong sort of robots—while simultaneously failing to show that some robots are not conscious.)

Pro-civil-libertarian Arguments

If the usual "anti-civil-libertarian" arguments (arguments against conceding that Oscar is conscious) are bad arguments, *pro*-civil-libertarian arguments seem to be just about nonexistent! Since the nineteenth century, materialists have contended that "consciousness is just a property of matter at a certain stage of organization." But as a semantic analysis this contention is hopeless (psychophysical parallelism is certainly not *analytic*), and as an identity theory it is irrelevant. Suppose that Feigl had been correct, and that sensation words *referred* to events (or "states" or "processes") definable in the language of physics. (As I remarked before, Feigl no longer holds this view.) In particular, suppose 'the sensation of red' *denotes* a brain process. (It is, of course, utterly unclear what this supposition comes to. We are taught the use of 'denotes' in philosophy by being told that 'cat' denotes the class of all cats, and so on; and then some philosophers say " 'the sensation of red' denotes a class of brain processes," as if *this* were now supposed to be clear! In fact, all we have been told is that " 'the sensation of red' denotes a brain process" is true just in case "the sensation of red *is* a brain process" is

true. Since this latter puzzling assertion was in turn explained by the identity theorists in terms of the distinction between *denotation* and *connotation,* nothing has been explained.) Still, this does not show that Oscar is conscious. Indeed, Oscar may be psychologically isomorphic to a human without being at all similar in physical-chemical construction. So we may suppose that Oscar does not have "brain processes" at all and, hence, (on this theory) that Oscar is *not* conscious. Moreover, if the physical "correlate" of the sensation of red (in the case of a human) is P_1, and the physical correlate of the "sensation" of red (in the case of Oscar) is P_2, and if P_1 and P_2 are *different* physical states, it can nonetheless be maintained that, when Oscar and I both "see something that looks red" (or "have the sensation of red," to use the philosophical jargon that I have allowed myself in this paper), we are in the *same* physical state, namely the *disjunction* of P_1 and P_2. How do we decide whether "the sensation of red" (in the case of a human) is "identical" with P_1 or "identical" with $P_1 \vee P_2$? Identity theorists do not tell me anything that helps me to decide.

Another popular theory is that ordinary-language psychological terms, such as 'is angry' (and, presumably, such quasi-technical expressions as 'has the sensation of red') are *implicitly defined by a psychological theory.* On this view, it would follow from the fact that Oscar and I are "models" of the same psychological (molar behavioral) theory that psychological terms have *exactly the same sense* when applied to me and when applied to Oscar.

It may, perhaps, be granted that there is something that could be called an "implicit psychological theory" underlying the ordinary use of psychological terms. (That an angry man will behave aggressively, unless he has strong reasons to repress his anger and some skill at controlling his feelings; that insults tend to provoke anger; that most people are not very good at controlling strong feelings of anger; are examples of what might be considered "postulates" of such a theory. Although each of these "postulates" is quasi-tautological, it might be contended that the conjunction of a sufficient number of them has empirical consequences, and can be used to provide empirical explanations of observed behavior.) But the view that the whole meaning of such a term as 'anger' is fixed by its place in such a theory seems highly dubious. There is not space in the present paper to examine this view at the length that it deserves. But one or two criticisms may indicate where difficulties lie.

To assert that something contains phlogiston is (implicitly) to

assert that certain laws, upon which the concept of phlogiston depends, are correct. To assert that something is electrically charged is in part to assert that the experimental laws upon which the concept of electricity is based and which electrical theory is supposed to explain, are not radically and wholly false. If the "theory" upon which the term anger "depends" really has empirical consequences, then even to say "I am angry" is in part to assert that these empirical consequences are not radically and wholly false. Thus it would not be absurd, if 'anger' really *were* a theoretical term, to say "I think that I am very angry, but I'm not sure" or "I think that I have a severe pain, but I'm not sure" or "I think that I am conscious but I'm not sure," since one might well not be sure that the experimental laws implied by the "psychological theory" implicit in ordinary language are in fact correct. It would also not be absurd to say: "perhaps there is not really any such thing as anger" or "perhaps there is not really any such thing as pain" or "perhaps there is not really any such thing as being conscious." Indeed, no matter how certain I might be that I have the sensation of red, it might be proved *by examining other people* that I did *not* have that sensation and that in fact there was no such thing as having the sensation of red. Indeed, "that *looks like* the sensation of red" would have a perfectly good use—namely, to mean that my experience is as it would be if the "psychological theory implicit in ordinary language" were true, but the theory is not in fact true. These consequences should certainly cast doubt on the idea that "psychological terms in ordinary language" really are "theoretical constructs."

It is obvious that "psychological terms in ordinary language" have a *reporting use*. In the jargon of philosophers of science, they figure in *observation statements*. "I am in pain" would be such a statement. But clearly, a term that figures in observational reports has an observational use, and that use must enter into its meaning. Its meaning cannot be fixed merely by its relation to other terms, in abstraction from the actual speech habits of speakers (including the habits upon which the reporting use depends).

The first difficulty suggests that the "psychological theory" that "implicitly defines" such words as 'anger' has in fact *no* nontautological consequences—or, at least, no empirical consequences that could not be abandoned without changing the meaning of these words. The second difficulty then further suggests that the job of fixing the meaning of these words is only partially done by the logical relationships (the "theory"), and is completed by the reporting use.

A third difficulty arises when we ask just what it is that the "psychological theory implicit in ordinary language" is supposed to be *postulating*. The usual answer is that the theory postulates the existence of certain *states* which are supposed to be related to one another and to behavior as specified in the theory. But what does 'state' mean? If 'state' is taken to mean physical state, in the narrow sense alluded to before, then psychophysical parallelism would be implied by an arbitrary "psychological" assertion, which is obviously incorrect. On the other hand, if 'state' is taken in a sufficiently wide sense so as to avoid this sort of objection, then (as Wittgenstein points out) the remark that "being angry is being in a certain psychological state" *says nothing whatsoever.*

In the case of an ordinary scientific theory (say, a physical theory), to postulate the existence of "states" S_1, S_2, . . . , S_n satisfying certain postulates is to assert that one of two things is the case: either (1) physical states (definable in terms of the existing primitives of physical theory) can be found satisfying the postulates; or (2) it is necessary to take the new predicates S_1, . . . , S_n (or predicates in terms of which they can be defined) as additional primitives in physical science, and widen our concept of "physical state" accordingly. In the same way, identity theorists have sometimes suggested that "molar psychological theory" *leaves it open* whether or not the states it postulates are physical states or not. But if physical states *can* be found satisfying the postulates, then they are the ones referred to by the postulates. 'State' is then a methodological term, so to speak, whose status is explained by a perspicuous representation of the procedures of empirical theory construction and confirmation. This solution to our third difficulty reduces to the identity theory under the supposition that psychophysical parallelism holds, and that physical states *can* be found "satisfying" the postulates of "molar behavioral psychology."

Even if this solution to the third difficulty is accepted, however, the first two difficulties remain. To be an empirically confirmable scientific theory, the "molar behavioral theory" implicit in the ordinary use of psychological terms must have testable empirical consequences. If the ordinary-language psychological terms really designate states postulated by this theory, then, if the theory is radically false, we must say there are no such "states" as being angry, being in pain, having a sensation, etc. And this must always remain a possibility (on this account), no matter what we observe, since no finite number of observations can deductively establish a scientific theory properly so-called. Also, the report-

ing role of "psychological" terms in ordinary language is not discussed by this account. If saying "I am in pain" is simply ascribing a *theoretical* term to myself, then this report is in part a *hypothesis,* and one which may always be false. This account —that the ordinary use of "psychological" terms presupposes an empirical theory, and one which may be radically false—has recently been urged by Paul Feyerabend. Feyerabend would accept the consequence that I have rejected as counterintuitive: that there may not really be any pains, sensations, etc., in the customary sense. But where is this empirical theory that is presupposed by the ordinary use of "psychological" terms? Can anyone state *one* behavioral law which is clearly empirical and which is presupposed by the concepts of sensation, anger, etc.? The empirical connection that exists, say, between being in pain and saying "ouch," or some such thing, has sometimes been taken (by logical behaviorists, rather than by identity theorists) to be such a law. I have tried to show elsewhere,[4] however, that no such law is really required to be true for the application of the concept of pain in its customary sense. What entitles us to say that a man is in pain in our world may not entitle one to say that he is in pain in a different world; yet the *same* concept of pain may be applicable. What I contend is that to understand any "psychological" term, one must be implicitly familiar with a network of *logical* relationships, and one must be adequately trained in the reporting use of that word. It is also necessary, I believe, that one be prepared to accept first-person statements by other members of one's linguistic community involving these predicates, at least when there is no *special* reason to distrust them; but this is a general convention associated with discourse, and not part of the meaning of any particular word, psychological or otherwise. Other general conventions associated with discourse, in my opinion, are the acceptance of not-too-bizarre rules of inductive inference and theory confirmation and of certain fundamental rules of deductive inference. But these things, again, have to do with one's discourse *as a whole* not being linguistically deviant, rather than with one's understanding any particular word. If I am not aware that someone's crying out (in a certain kind of context) is a sign that he is in pain, I can be *told.* If I refuse (without good reason), to believe what I am told, it can be pointed out to me that, when I am in that context (say, my finger is burnt), *I* feel pain, and no

[4] In "Brains and Behavior." The character of psychological concepts is also discussed by me in "The Mental Life of Some Machines," to appear in a forthcoming collection edited by Hector Neri Castañeda.

condition known by me to be relevant to the feeling or nonfeeling of pain is different in the case of the Other. If I *still* feel no inclination to ascribe pain to the Other, then my whole concept of discourse is abnormal—but it would be both a gross understatement and a misdiagnosis to say that I "don't know the meaning of 'pain'."

I conclude that "psychological" terms in ordinary language are *not* theoretical terms. Moreover, the idea that, if psychophysical parallelism is correct, then it is analytic that pain *is* the correlated brain-state is not supported by a shred of linguistic evidence. (Yet this is a consequence of the combined "identity theory–theoretical term" account as we developed it to meet our third difficulty.) I conclude that any attempt to show that Oscar is conscious (analytically, relative to our premises) along these lines is hopeless.

Ziff's Argument

So far all the arguments we have considered, on both sides of the question: Is Oscar conscious?, have been without merit. No sound consideration has been advanced to show that it is false, given the meaning of the words in English and the empirical facts as we are assuming them, that Oscar is conscious; but also no sound consideration has been advanced to show that it is true. If it is a violation of the rules of English to say (without "raised-eyebrow quotes") that Oscar is in pain or seeing a rose or thinking about Vienna, we have not been told *what* rules it violates; and if it is a violation of the rules of English to *deny* that Oscar is conscious, given his psychological isomorphism to a human being, we have likewise not been told what rules it violates. In this situation, it is of interest to turn to an ingenious ("anti-civil-libertarian") argument by Paul Ziff.[6]

Ziff wishes to show that it is false that Oscar is conscious. He begins with the undoubted fact that if Oscar is not alive he cannot be conscious. Thus, given the semantical connection between 'alive' and 'conscious' in English, it is enough to show that Oscar is not *alive*. Now, Ziff argues, when we wish to tell whether or not something is alive, we do *not* go by its *behavior*. Even if a thing looks like a flower, grows in my garden like a flower, etc., if I find upon taking it apart that it consists of gears and wheels

[6] I take the liberty of reporting an argument used by Ziff in a conversation. I do not wish to imply that Ziff necessarily subscribes to the argument in the form in which I report it, but I include it because of its ingenuity and interest.

and miniaturized furnaces and vacuum tubes and so on, I say "what a clever mechanism," not "what an unusual plant." It is *structure,* not *behavior* that determines whether or not something is alive; and it is a violation of the semantical rules of our language to say of anything that is clearly a mechanism that it is "alive."

Ziff's argument is unexpected, because of the great concentration in the debate up to now upon *behavior,* but it certainly calls attention to relevant logical and semantical relationships. Yet I cannot agree that these relationships are as clear-cut as Ziff's argument requires. Suppose that we construct a robot—or, let me rather say, an *android,* to employ a word that smacks less of mechanism—out of "soft" (protoplasm-like) stuff. Then, on Ziff's account, it may be perfectly correct, if the android is sufficiently "life-like" in structure, to say that we have "synthesized life." So, given two artifacts, both "models" of the same psychological theory, both completely deterministic physical-chemical systems, both designed to the same end and "programmed" by the designer to the same extent, it may be that we must say that one of them is a "machine" and not conscious, and the other is a "living thing" (albeit "artificially created") and conscious, simply because the one consists of "soft stuff" and the other consists of "hardware." A great many speakers of English, I am sure (and I am one of them), would find the claim that this dogmatic decision is required by the meaning of the word 'alive' quite contrary to their linguistic intuitions. I think that the difficulty is fundamentally this: a plant does not exhibit much "behavior." Thus it is natural that criteria having to do with *structure* should dominate criteria having to do with "behavior" when the question is whether or not something that looks and "behaves" like a plant is really a living thing or not. But in the case of something that looks and behaves like an *animal* (and especially like a *human being*), it is natural that criteria having to do with behavior—and not just with actual behavior, but with the *organization* of behavior, as specified by a psychological theory of the thing—should play a much larger role in the decision. Thus it is not unnatural that we should be prepared to argue, in the case of the "pseudo-plant," that "it isn't a living thing because it is a mechanism," while some are prepared to argue, in the case of the robot, that "it isn't a *mere* mechanism, because it is *alive,*" and "it is alive, because it is conscious," and "it is conscious because it has the same behavioral organization as a living human being." Yet Ziff's account may well explain why it is that many speakers are not convinced by these latter arguments. The ten-

sion between conflicting criteria results in the "obviousness," to some minds, of the robot's "machine" status, and the equal "obviousness," to other minds, of its "artificial-life" status.

There is a sense of 'mechanism' in which it is clearly analytic that a mechanism cannot be alive. Ziff's argument can be reduced to the contention that, on the normal interpretation of the terms, it is analytic in English that something whose *parts* are all mechanisms, in this sense, likewise cannot be alive. If this is so, then no English speaker should suppose that he could even *imagine* a robot *thinking,* being *power-mad, hating humans,* or *being in love,* any more than he should suppose that he could imagine a married bachelor. It seems evident to me (and indeed to most speakers) that, absurdly or not, we *can* imagine these things. I conclude, therefore, that Ziff is wrong: it may be *false,* but it is not a *contradiction,* to assert that Oscar is alive.

The "Know-Nothing" View

We have still to consider the most traditional view on our question. According to this view, which is still quite widely held, *it is possible that Oscar is conscious, and it is possible that he is not conscious.* In its theological form, the argument runs as follows: I am a creature with a body and a soul. My body happens to consist of flesh and blood, but it might just as well have been a machine, had God chosen. Each voluntary movement of my body is correlated with an activity of my soul (how and why is a "mystery"). So, it is quite possible that Oscar has a soul, and that each "voluntary" movement of his mechanical body is correlated in the same mysterious way with an activity of his soul. It is also possible—since the laws of physics suffice to explain the motions of Oscar's body, without use of the assumption that he has a soul —that Oscar is but a lifeless machine. There is absolutely no way in which we can know. This argument can also be given a nontheological (or at least apparently nontheological) form by deleting the reference to God, and putting 'mind' for 'soul' throughout. To complete the argument, it is contended that I know what it *means* to say that Oscar has a "soul" (or has a pain, or the sensation of red, etc.) *from my own case.*

One well-known difficulty with this traditional view is that it implies that it is also possible that other humans are not really conscious, even if they are physically and psychologically isomorphic to me. It is contended that I can know with *probability* that other humans are conscious by the "argument from analogy." But in the inductive sciences, an argument from analogy is gen-

erally regarded as quite weak unless the conclusion is capable of further and independent inductive verification. So it is hard to believe that our reasons for believing that other persons are conscious are very strong ones if they amount simply to an analogical argument with a conclusion that admits of *no* independent check, observational, inductive, or whatever. Most philosophers have recently found it impossible to believe *either* that our reasons for believing that other persons are conscious are that weak *or* that the possibility exists that other persons, while being admittedly physically and psychologically isomorphic (in the sense of the present paper) to myself, are not conscious. Arguments on this point may be found in the writings of all the major analytical philosophers of the present century. Unfortunately, many of these arguments depend upon quite dubious theories of meaning.

The critical claim is the claim that it follows from the fact that I have had the sensation of red, I can imagine this sensation, I "know what it is like," that I can understand the assertion that Oscar has the sensation of red (or any other sensation or psychological state). In a sense, this is right. I *can,* in one sense, understand the *words.* I can parse them; I don't think "sensation of red" means *baby carriage,* etc. More than that: I know what I would experience if I were conscious and psychologically as I am, but with Oscar's mechanical "body" in place of my own. How does this come to be so? It comes to be so, at least in part, because we have to learn from experience what our own bodies are like. If a child were brought up in a suitable kind of armor, the child might be deceived into thinking that it was a robot. It would be harder to fool him into thinking that he had the internal structure of a robot, but this too could be done (fake X rays, etc.). And when I "imagine myself in the shoes of a (conscious) robot," what I do, of course, is to imagine the sensations that I might have if I were a robot, or rather *if I were a human who mistakenly thought that he was a robot.* (I look down at my feet and see bright metal, etc.)

Well, let us grant that in this sense we *understand* the sentence "Oscar is having the sensation of red." It does not follow that the sentence possesses a truth value. We understand the sentence "the present King of France is bald," but, on its normal interpretation in English, the sentence has no truth value under present conditions. We can give it one by adopting a suitable convention—for example, Russell's theory of descriptions—and more than one such suitable convention exists. The question really at issue is *not* whether we can "understand" the sentences "Oscar

is conscious'' (or ''has the sensation of red'' or ''is angry'') and ''Oscar is not conscious,'' in the sense of being able to use them in such contexts as ''I can perfectly well picture to myself that Oscar is conscious,'' but whether there really is an intelligible sense in which one of these sentences is true, on a normal interpretation, and the other false (and, in that case, whether it is also true that we can't tell which).

Let us revert, for a moment, to our earlier fantasy of ROBOTS —i.e., second-order robots, robots created by robots and regarded by the robots as *mere* ROBOTS. As already remarked, a robot philosopher might very well be led to consider the question: Are ROBOTS conscious? The robot philosopher ''knows,'' of course, just what ''experiences'' he would have if he were a ''conscious'' ROBOT (or a robot in a ROBOT suit). He can ''perfectly well picture to himself that a ROBOT could have ''sensation.'' So he may perfectly well arrive at the position that it is logically possible that ROBOTS have sensations (or, rather, ''sensations'') and perfectly possible that they do not, and moreover he can never know. What do we think of this conclusion?

It is clear what we should think: we should think that there is not the slightest reason to suppose (and every reason not to suppose) that there is a special property, ''having the 'sensation' of red,'' which the ROBOT may or may not have, but which is inaccessible to the robot. The robot, knowing the physical and psychological description of the ROBOT, is in a perfectly good position to answer all questions about the ROBOT that may reasonably be asked. The idea that there is a further question (class of questions) about the ROBOT which the robot cannot answer, is suggested to the robot by the fact that these alleged ''questions'' are grammatically well formed, can be ''understood'' in the sense discussed above, and that the possible ''answers'' can be ''imagined.''

I suggest that our position with respect to robots is *exactly* that of robots with respect to ROBOTS. There is not the slightest reason for us, either, to believe that ''consciousness'' is a well-defined property, which each robot either *has* or *lacks,* but such that it is not possible, on the basis of the physical description of the robot, or even on the basis of the psychological description (in the sense of ''psychological'' explained above), to *decide* which (if any) of the robots possess this property and which (if any) fail to possess it. The rules of ''robot language'' may well be such that it is perfectly possible for a robot to ''conjecture'' that ROBOTS have ''sensations'' and also perfectly possible for a robot

to conjecture that ROBOTS do not have "sensations." It does not follow that the physical and psychological description of the ROBOTS is "incomplete," but only that the concept of "sensation" (in "raised-eyebrow quotes") is a well-defined concept only when applied to robots. The question raised by the robot philosopher: Are ROBOTS "conscious"? calls for a decision and not for a discovery. The decision, at bottom, is this: Do I treat ROBOTS as fellow members of my linguistic community, or as machines? If the ROBOTS are accepted as full members of the robot community, then a robot can find out whether a ROBOT is "conscious" or "unconscious," "alive" or "dead" in just the way be finds out these things about a fellow robot. If they are rejected, then nothing *counts* as a ROBOT being "conscious" or "alive." Until the decision is made, the statement that ROBOTS are "conscious" has no truth value. In the same way, I suggest, the question: Are robots conscious? calls for a decision, on our part, to treat robots as fellow members of our linguistic community, or not to so treat them. As long as we leave this decision unmade, the statement that robots (of the kind described) are conscious has no truth value.

If we reject the idea that the physical and psychological description of the robots is incomplete (because it "fails to specify whether or not they are conscious"), we are not thereby forced to hold either that "consciousness" is a "physical" attribute or that it is an attribute "implicitly defined by a psychological theory." Russell's question in the philosophy of mathematics: If the number 2 is not the set of all pairs, then what on earth is it? was a silly question. Two is simply the second number, and nothing else. Likewise, the materialist question: If the attribute of "consciousness" is not a physical attribute (or an attribute implicitly defined by a psychological theory) then what on earth is it? is a silly question. Our psychological concepts in ordinary language are as we have fashioned them. The "framework" of ordinary-language psychological predicates is what it is and not another framework. *Of course* materialism is false; but it is so *trivially* false that no materialist should be bothered!

Conclusion

In this paper, I have reviewed a succession of failures: failures to show that we *must* say that robots are conscious, failures to show that we *must* say they are not, failures to show that we *must* say that we can't tell. I have concluded from these failures that there is no correct answer to the question: Is Oscar conscious? Robots

may indeed have (or lack) properties unknown to physics and undetectable by us; but not the slightest reason has been offered to show that they do, as the ROBOT analogy demonstrates. It is reasonable, then, to conclude that the question that titles this paper calls for a decision and not for a discovery. If we are to make a decision, it seems preferable to me to extend our concept so that robots *are* conscious—for "discrimination" based on the "softness" or "hardness" of the body parts of a synthetic "organism" seems as silly as discriminatory treatment of humans on the basis of skin color. But my purpose in this paper has not been to improve our concepts, but to find out what they are.

HILARY PUTMAN

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MERE ROBOTS AND OTHERS

"MY attitude towards [another human being] . . . is an attitude towards a soul. I am not of the *opinion* that he has a soul" (*Philosophical Investigations* II, iv). Something like that is right, I think, though I have misgivings about it. If it is right, then so is what I take to be the principal moral of Putnam's paper for the stock problem of other minds. But I think he is too ready to give robots the vote.

To begin with, the notion of a robot "psychologically isomorphic" to us, in Putnam's sense, presupposes a good deal about *us* that seems to me doubtful or at least not obviously true. And although Putnam assumes for the sake of argument the truth of what he calls "psychophysical parallelism," I am inclined to argue that this assumption is neither clear nor plausible. But suppose that it is both, and is true, and (in addition) that our bodies are deterministic systems. And suppose that some deterministic robot is psychologically isomorphic to us and "talks" incessantly. Shall we give in and accept it, in Putnam's phrase, as a fellow member of our linguistic community? I don't doubt that in an actual case we might helplessly do that, whatever it amounts to; and if we did, the question whether the new member

* Abstract of a paper to be presented in a symposium on "Minds and Machines" at the sixty-first annual meeting of the American Philosophical Association, Eastern Division, December 28, 1964, commenting on Hilary Putnam, "Robots: Machines or Artificially Created Life?", this JOURNAL, **61**, 21 (Nov. 12, 1964): 668–691.