Witold Hensel

From Weak to Strong Functionalism, and Back

Although it is customary to begin the discussion of functionalism with an introduction of the notion of Turing machines, such a construal of it, engaging the apparatus of mathematics at the level of the doctrine's most fundamental concepts, has the disadvantage of obscuring some of its aspects which deserve highlighting before we proceed with a more detailed analysis. It is well to remark at the outset that the very idea of the mind's being a functionally organized structure can easily be formulated without recourse to a particular theory of computability or, indeed, without any mention of the computer metaphor. What seems to underlie all the technical jargon of various functionalist avowals, whether they employ the language of neural networks or the formalism of Ramsey sentences, is the basic intuition that mental states are in fact or should in principle be defined on the basis of causal connections which relate them to other mental states, on the one side, and to stimuli and responses, on the other.

One virtue of this informal definition is that it shows that functionalism springs from the two apparently conflicting sources we have discussed above, namely identity theory and logical behaviorism; it combines the behaviorist insight that mental states are individuated by virtue of their causes and effects with the physicalist claim that they are situated in the brain and cannot be reduced to behavioral dispositions. It is actually very tempting to see in functionalism a synthesis, as it were, of these two broadly materialist views. And my use of Hegelian dialectics in this connection is not merely stylistic either; it points to a disconcerting fact about all syntheses, including functionalism: a combination of two good ideas may not be such a good idea, after all. Just as the proponents of functionalism will argue that their doctrine takes the best of both philosophical worlds, so its opponents will claim that it is open to all the devastating charges made against behaviorism and identity theory, and possibly some new ones. At any rate, our final verdict will depend on some pretty general philosophical convictions.

Another virtue of my loose construal of the functionalist thesis is that it does not commit the functionalist to any particular notion of computability or research program in AI, thereby accommodating for future developments in mathematics and computer science. It is true that by 1960, when Putnam first suggested the position, Church's thesis, asserting the equivalence of the intuitive notion of what a human being can in principle calculate with Turing-machine computability, had been so firmly established that its employment in a philosophical theory must have been entirely uncontroversial. However, the advent of connectionism and other mathematical results have since made the agreement on that point less general. By my informal definition, connectionism is certainly and unambiguously a variety of functionalism, which in turn reveals its functionalist roots, logical and historical.[1]

The advantage of generality in philosophy is often accompanied by a serious deficiency in content, however; and my definition of functionalism is no exception to this piece of professional wisdom. It goes without saying that all the crucial terms present in it will remain almost completely meaningless if no theory is supplied which explicates the notions of state and causal connection. The theory employed by Putnam is, of course, the theory of Turing machines.

As is well-known, a Turing machine is an abstract, mathematical description of the computer. It consists of a tape, a scanner and a printing device. The tape may be finite or infinite, but what is now relevant is that it must be divided into squares, each of which may be read by the scanner detecting only the symbols of a specified finite alphabet. The printing device can fill the tape with letters, one square at a time; it can also move to the left and to the right or stay over the same square (L, R and C, respectively). The states of the machine fall into the following classes: a finite number of active states (S_1, S_2, …, S_n), the first of which is called the initial state, and one rest state (S_0), which the machine assumes upon stopping.

The crucial feature of a Turing machine is that its behavior is completely specified by the program it follows and by the contents of the input-output tape. The program may be presented in the form of a machine table, whose columns correspond to the machine's primitive symbols, while rows, to its internal states. A very simple machine table, taken from Kleene (1952: 358), will suffice to show what such a description looks like. Machine M has one active state S_1 and one rest state S_0, and it is defined by the following table:

0	**1**
1 C S_0	**1** R S_1

If M is started over a square with **1** printed in it, it will print **1** and move to the right passing again into state S_1; if it encounters a **0** on the tape, it prints **1** in its place and assumes the rest state S_0.

§1. Weak Functionalism

Such Turing machines made their first appearance in Putnam 1960, ushering in the phase of functionalism in which Putnam used the computer analogy in order to prove that the whole philosophy of mind was confused and irrelevant.[2] Interestingly enough, he attempted to achieve

[1] Interestingly, Putnam (1964: 394) writes: '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 neurons, would have exactly the same psychology as a human.'

[2] Putnam 1960 tackled the mind-body problem, Putnam 1963 aimed to bury logical behaviorism, Putnam 1964 dealt with the question of other minds.

this without making any positive claims about the nature of mind; the only theses he asserted seem to have been: (1) if all philosophical problems have their logical analogues in terms of Turing machines, then they have nothing to do with our makeup[3] (Putnam's additional contention is that the antecedent of this conditional holds), and (2) a review of standard arguments in the philosophy of mind reveals that the issues involved call for a decision rather than a discovery.

Let us call this phase *weak functionalism* to distinguish it from the stronger claim made in later papers. To see what this weak functionalism looked like, let us construct the analogy and trace a selection of philosophical arguments couched in terms of Turing machines. What we need is a community of robots, that is finite automata (realizations of Turing machines with a finite tape) with some kind of electronic sense organs and the capacity to move. Furthermore, their internal architecture should allow them to access information about some of their internal states, enabling them to issue 'first-person' reports of their inner workings (something along the lines of malfunction alerts).

The possibility of mapping problems in analytic philosophy onto finite automata depends not only on the architecture of the robots but also on the state of their knowledge; thus, we stipulate that they be as scientifically and philosophically advanced as people were in the late 1950's. More specifically, they can construct empirical theories in accordance with all the known principles of methodology, but they do not possess complete knowledge of their internal structure or of the outside world; they hold that scientific theories are formal calculi which are partially interpreted by means of correspondence rules.

Now suppose the body of knowledge of our robots contains a number of correlation statements such as 'Whenever I am in state A, flip-flop 105 is on.' Since 'I am in state A.' is an observational statement and 'Flip-flop 105 is on' is a theoretical one, a controversy may ensue among robot philosophers as to whether or not correlations of the occurrence of internal states with flip-flops being on or off may be explained by asserting an identity between the property of being in state A and the property of flip-flop 105 being on. (Such an identification would have to construe stimuli and responses as purely physical happenings to avoid a category mistake; so, for example, an utterance would be described as a series of sounds rather than a message.)

Some robots may object to such an identification on the basis that identity statements are acceptable only if they are analytic: since any statement of the form 'State … is identical with flip-flop … being on/off' must be synthetic, no identity can possibly obtain between internal states and flip-flops. According to Putnam, this charge presupposes two questionable premises: (1) that

[3] This is an overstatement: obviously, not every function can be realized by any sort of material.

there is a clear distinction between analytic and synthetic statements, and (2) that properties are identical with meanings.

Indeed, other robots may point out that the first premise has pretty much been refuted by a Quine-1950. Anyhow, the premises jointly imply that no identifying reductions in empirical science are possible. The identification of light, they will say, with electromagnetic radiation of certain wavelengths is not analytic, yet we do not see it as illegitimate; any semantic theory which precludes such reductions is therefore inadequate.

Some robots, skilled in subtle analysis, may challenge this line of defense by drawing attention to the difference between properties, on the one hand, and events and objects, on the other: it is true that we can describe one event or object in two or more nonequivalent ways (reduction of light to electromagnetic radiation is an example of identifying events, the identification of water with H_2O can be viewed as a reduction of objects), but it does not follow from this that we may proceed in a similar fashion with properties. Indeed, some semantic theories explicitly identify meanings with properties (Frege, Carnap, Lewis), so, by definition, if two expressions differ in meaning, they signify different properties.

This only prompts the materialist robots to offer more examples of reductions: the fluidity of liquid being identical with its 'loose' atomic structure, or the property of being red being the same as a disposition to reflect electromagnetic radiation of certain wavelengths. Alternatively, they may argue for a better theory of meaning. As for Putnam, he observes that if a human philosopher employs this argument against identification, he must be prepared to 'hug the souls of Turing machines to his philosophical bosom!' (1960: 376) because the sentences 'I am in state A' and 'Flip-flop 105 is on' are not synonymous in the robots' language by any standard – their methods of verification are different, and so are their uses.

Another argument against the identification of internal states with flip-flops which the robots may consider is of a linguistic character. *The linguistic argument* asserts that the sentence 'Being in state A is identical with having flip-flop 105 on' is deviant in the sense that there is no statement which it can be used to express in a normal context. Presumably, it cannot make sense unless it simply expresses a convention stipulating that the phrase 'state A' be henceforth replaced by the phrase 'flip-flop 105 is on'. But this would be tantamount to giving the words a new meaning, because, in accord with the new convention, the statement 'My flip-flop 105 is on' merely acquires the same meaning as 'I am in state A'; hence, advancing this kind of identity involves either a change of meaning or a grammatical error (in fact, both disjuncts are exceedingly difficult to prove).

A possible reply to this objection is that one should construe the problem diachronically rather than synchronically. Even if one accepts the charge that the identity statement at hand is deviant, it does not preclude it from becoming non-deviant in the future; so the real question is whether or not, and under what conditions, such an utterance could acquire a standard use. This remark would probably suffice to block the linguistic argument, since it draws attention to two crucial considerations, namely that: (a) language changes over time, and (b) predicting that some change in language will not take place would require precognition.

However, it is worth emphasizing some other points of disagreement between the standard Wittgensteinian approach to meaning, assumed by the anti-reductionist robot, and the one Putnam advocates, inspired by Ziff 1960:

(1) When a deviant sentence acquires a standard use, no change in meaning need be involved. For example, 'I am a hundred miles away from you' in ancient Greek would have been deviant prior to the invention and spread of writing. Yet, it is obvious that its' acquiring a standard use had nothing to do with a shift in meaning. For one thing, since the significance of a sentence is determined by the meaning of its parts and rules of composition, if one contends that a sentence changed in meaning then one must also be able to show which *words* or which rules of composition contributed to the change. But, clearly, none did. What transpires is that the new use was far from arbitrary, 'but,' as Putnam (1960: 378) puts it, 'represented an automatic projection from the existing stock uses of the several words making up the sentence, given the new context'. Just as new technology may bring about new sentence uses, so can advances in theory; thus, an identification of flip-flop states with internal states, or pains with the stimulation of C-fibers, may become non-deviant in a suitable theoretic context.

(2) It seems that the 'is' occurring in the sentence 'State A is identical with flip-flop 105 being on' expresses a theoretical identification. Such an identification is perfectly intelligible, permissible and even advisable when it (a) allows us to derive to a good approximation the laws of a reduced theory from the laws of a reducing theory, (b) enables us to make new predictions, (c) explains the failures of the reduced theory, (d) yields a fruitful research program. It also leads to increased ontological simplicity and serves to reject certain questions as nonsensical (for example, 'What is pain over and above being the stimulation of C-fibers?').

(3) A theoretical identification does not normally allow us to replace a reduced term by a reducing one in every context; it is still deviant to say 'Pass me some H_2O', even if we all agree that water *is* H_2O.

It follows from these considerations that the linguistic argument assumes a false conception of meaning, by Putnam's lights, but even if it did not, the robot identity theorist may still argue that he is essentially making a claim about future developments in science – *ergo* the linguistic argument misses his point. On the other hand, the skeptically minded robot can now object that his opponent cannot justify his position without a suitable reducing theory in place – I think this is not as strong a reason for rejecting reductionist doctrines as it is often taken to be, but it seems that Putnam, at this point, is impressed; I will return to this in §2. (There is one essential point to be made in this connection: if one is to make the positive functionalist claim that mental states are identical with the states of Turing machines then one is *ipso facto* committed to the permissibility of inter-theoretic identifications.)

So far we have discussed the mind-body problem as translated into 'Robotese'; we saw that essentially all relevant aspects of the debate arise also for our society of Turing machines of the 1950's. Assuming that our robot community is unaware of the existence of people, we observe that the identity theorist had made a strong case for reduction, even if there were some semantic difficulties that could not be dispelled. Taking a step back, however, we cannot help noticing that the robot's physicalist answer to the mind-body problem is irrelevant in the sense that it excludes a number of possible systems which happen to be composed of something other than flip-flops (namely, us – humans).

Now we can ask ourselves two fundamental questions:

(i) What are the conditions which must be satisfied for the robot analogy to work?

(ii) Are they really satisfied?

Putnam's reply to (i) is surprisingly simple: he asserts that the mind-body puzzles arise for any system that uses language, does not have full knowledge of its physical makeup, and comes to know its physical structure by constructing theories and experimenting.

This answer may be criticized as question-begging because it assumes at the outset that robots have a mental life: only systems possessing mental states, it seems, can be said to have knowledge and strive to extend it through theorizing and experimentation.

But how can a robot really know anything? Isn't there an aspect of the mind which makes the analogy ineffective? Clearly, before accepting Putnam's analogy, we must deal with the problem of other minds. Two standard objections to describing robots as 'persons' are as follows:

> (1) Some *qualia* have the intrinsic property of being pleasurable or painful; a robot cannot have such *qualia* because it can be reprogrammed so that what used to be pleasurable will be painful, or vice versa.

(2) When a person utters the sentence ‘I see red’, he usually does so because he *knows* that he is having a sensation of seeing red. A robot does not know, however; it is simply *caused* to generate the appropriate sound.

As to (1), what this objection overlooks is that a robot may be so complex that no reprogramming would be possible. Secondly, it may be argued that human *qualia* are not intrinsically pleasurable or painful: any acquired taste can be said to produce *qualia* which become pleasurable only after sufficient ‘training’.

As to (2), there are a number of different interpretations of what it means ‘to know that one has a sensation’ – all of them are compatible with a sufficiently robust robot model. This objection works only against the so-called *evincing model* described in Putnam 1960; that model, however, was inspired by Wittgenstein, who thought that sensations are not objects of knowledge.[4] There are two crucial assumptions Putnam makes here, which need to be highlighted. The first assumption, known as *psychological isomorphism*, embodies functionalism on my loose construal and says that, for whatever psychological theory we may devise, there may exist non-human (possibly artificial) systems which obey the same psychological laws as we do. If psychological isomorphism does not hold, the robot analogy will be broken. The second assumption is *conceptual-role semantics*: the terms of a theory are implicitly defined by the laws of the theory. *Ergo*, if our robots fall in the domain of true psychology, then they possess mental states.

It is well to note in this connection how effectively Putnam utilizes both these assumptions. The strategy he adopts can be expressed as the following challenge: ‘If you think that you can disprove my position, then you must show me where exactly the robot analogy breaks down. To do that you must produce a psychological theory which is true of humans but false of my robots, for, otherwise, your argument will lack content. If you should succeed in providing such a theory, however, then I will be able to construct a robot model for it!’ This game can be played *ad nauseam*, the result always being the same...

The only way out is to attack head on at least one of Putnam’s assumptions. As to conceptual-role semantics, it may be argued, for instance, that an accurate account of meaning entails that no mental term applies to robots, even if they are psychologically isomorphic to humans. Further discussion of this possibility would transport us into the philosophy of language, which lies outside the scope of this paper. Let me, for the moment, accept both psychological isomorphism and conceptual-role semantics without a quibble and answer question (ii) in the affirmative.

[4] The evincing model was constructed so as to issue a pain report immediately on detecting a certain state. See Putnam (1960: 368) for details.

At this point Putnam, somewhat surprisingly, proposes that we see whether there are any reasons for denying consciousness to robots while, at the same time, ascribing it to humans. In order not to beg the question, he says, we must suppose that consciousness is not a mental state but a separate feature, which may or may not obtain of robots which are psychologically isomorphic to us – otherwise, the answer would be trivial. (Let me call this *the objectivity constraint.*)

Putnam discusses three arguments against awarding consciousness to such systems, and they are actually more humorous than convincing:

(1) *The phonograph-record argument* claims that the behavior of a robot is only played as if it were a phonograph record – the robot blindly follows a, possibly brilliant, set of instructions, written by the programmer. The fault of this argument is that it does not allow for learning robots. It goes without saying that only a learning robot can be psychologically isomorphic to us.

(2) *The reprogramming argument* asserts that a robot has no character of its own, as it can be reprogrammed at will. Barring the impossibility of reprogramming, we may point out that a human being can be reprogrammed as well, what with brain-washing, propaganda, medication and brain surgery.

(3) *The question-begging argument* states that human psychological states are not physical states, and it is not an argument at all.

All three arguments seem to hinge on the supposition that, since robots are artifacts, their apparently intelligent behavior must be a dim reflection of their designer's skill and imagination. To see how defective these charges are, all we have to do is imagine that we had discovered that we too were artifacts – would we immediately stop thinking of ourselves as persons?

The identity theorist would presumably argue that humans are conscious because they have brains, and robots are not because they do not, but, unless he can give some further reasons for this assertion, it will be sheer dogmatism. Paul Ziff, on the other hand, suggested that it is deviant to say of something that is not alive that it thinks – this, however, is as dogmatic as the physicalist dictum, for why should a bionic robot be more privileged than its aluminum counterpart?

Predictably, all arguments for awarding consciousness to robots point to psychological isomorphism and conceptual-role semantics (why not call the robot 'conscious' if it satisfies the laws of psychology to the same extent as any given human does?). It would have seemed that this is the most rational stance, especially if we see no grounds against taking it.

Putnam disagrees, asserting simply that arguments on both sides are, of necessity, inconclusive. More importantly, he is not being consistent! If mentality is a genuine phenomenon (that is, we adopt a realistic stance towards psychological theory) and psychological isomorphism holds, then Putnam's robots will have psychological states. However, on any reasonable construal of true psychology, consciousness must be a psychological state, which means that the objectivity constraint is misguided.

One way I can see to block the above reasoning would involve claiming that the blend of science, folk psychology and semantics at our disposal is not a theory mature enough to allow us to pass judgment on such potentially sensitive matters. After all, ascribing consciousness to a robot is tantamount to making a moral decision; if we really were faced with the question of whether or not we should grant civil rights a particular kind of machine, the importance of the moral dimension would become all too apparent. And this brings us to the second claim of weak functionalism (that the mind-body problem and the problem of other minds are in fact pseudo-problems – to repeat: they call for a decision rather than a discovery). What we have here is a *non sequitur*: from the fact that there might be a decision involved one cannot infer the stronger thesis that there is no room for rational discussion, even if the arguments on both sides have so far been inconclusive (which, on a certain level, is trivially true).

The issue of morality can be easily sidestepped, it seems, by adopting the conscious-until-proven-otherwise principle; in other words, the decision in question can be justified by appeal to ethical or pragmatic considerations. Putnam, however, has one more reply available: the choice of semantics is to a large extent arbitrary, and the arbitrariness of it renders traditional philosophy of mind irrelevant. Depending on what type of semantic theory we elect to assert, the problem of other minds and the mind-body problem will acquire a solution, one way or another.

There are, essentially, two things wrong with such a construal: no evidence is given in its support, and it runs counter to Putnam's most cherished philosophical views.

Be that as it may, let me summarize what we learned in this section. Firstly, we saw that it looks as though all philosophical problems might really have their robot analogues, which means that the assumption of psychological isomorphism is intuitive and might even be correct. Secondly, our discussion of the mind-body puzzle in terms of Robotese revealed that all of the arguments leveled against identity theory were unsound. Thirdly, it also showed that the traditional doctrines (monism *versus* dualism) had missed the simple point that the nature of mind might, at least to an extent, be independent of its particular realization. Thus, the first tenet of weak functionalism is warranted.

The second tenet, on the other hand, is unsubstantiated and inconsistent with the first. To repeat: if psychology is interpreted realistically, the claim that the problem of other minds has no solution is at odds with the thesis of psychological isomorphism, whose assertion is at the heart of the robot analogy. Unless Putnam is prepared to jettison scientific semantics, one consequence of this is that discoveries do force us to make certain kinds of decisions: if we have a well-confirmed theory that suggests that water is H_2O, it is reasonable to assert a suitable identity statement, and, conversely, it is unreasonable to refrain from asserting it.

The above leads us to abandon Putnam's reservations about the issue of other minds, and adopt strong functionalism.

A textual comment is in order here. In an autobiographical entry in Guttenplan (1994: 507), Putnam writes:

> In 1960 I published a paper titled 'Minds and Machines' which suggested a possible new option in the philosophy of mind, and in 1967 I published two papers ['The Mental Life of Some Machines' and 'The Nature of Mental States'] which became, for a time, the manifestos of the 'functionalist' current. … According to the functionalist view, a robot with the same program as a human being would *ipso facto* be conscious. Although in a talk to the American Philosophical Association in 1964 ['Robots: Machines of Artificially Created Life?'] I had drawn back from that view, … when I came to write the two papers I described as 'functionalist manifestos', I considered both the question as to whether psychological states are really 'functional' (i.e. computational) in nature and the question as to whether an automaton could be conscious to be factual questions. The earlier talk, I had come to see, contained an error.

This account suggests that Putnam asserted functionalism in 1960, and probably also in 1963, made the mistake of proposing its weak version in 1964, and then went back to the strong functionalist claim in 1967. In point of fact, however, the first functionalist paper begins with the following: 'The various issues and puzzles that make up the traditional mind-body problem are wholly linguistic and logical in character: whatever few empirical "facts" there may be in this area support one view as much as another.' (Putnam 1960: 362) The same point being reiterated in the conclusion: 'The moral, I believe, is quite clear: it is no longer possible to believe that the mind-body problem is a genuine theoretical problem, or that a "solution" to it would shed the slightest light on the world in which we live.' (1960: 384) Furthermore, in one of the 'functionalist manifestos' we find the following statements (1967a: 412): 'Today we know nothing strictly incompatible with the hypothesis that you and I are one and all Turing Machines, although we know some things that make this unlikely,' and, on page 424, 'As applied to Turing Machines, the functional organization is given by the machine table. A description of the functional organization of a human being might well be something quite and more complicated.' I have personally managed to find strong functionalism asserted only in 1967b and 1969. He criticized the position as early as 1973, in Putnam (1973c).

The above seems to suggest that Putnam was advocating weak functionalism from 1960 to 1967, and then asserted strong functionalism from 1967 to 1973.

§2. Strong Functionalism

In a nutshell, the strong functionalist's intuition is this: correct solutions to puzzles surrounding the mind are to be found by analyzing the structure of functionally organized systems; thus, the nature of consciousness will be discovered if we attend to a number of systems and find out what kind of functional structure is characteristic of those systems that we tend to call conscious ones. This clearly reductionist approach relies on two assumptions: (1) there is a set of laws which describe and explain the behavior of functional systems *qua* functional systems, and (2) either folk psychology is approximately true, that is, its concepts, such as the concept of consciousness, correspond to the essential functional features referred to in assumption (1), or we will arrive at a sufficiently robust psychological theory which may provide an empirical basis for the kind of theory mentioned in assumption (1).[5]

This way of expressing the research program of strong functionalism reveals that it must face a number of the same difficulties the identity theorist had to grapple with a while ago. For one thing, the strong functionalist needs to have two kinds of theories in place to get his project off the ground: a well-confirmed psychology, on the one hand, and a comprehensive functional systems theory, on the other. Secondly, since a reduction is viable only inasmuch as it explains and improves on the reduced theory, the functionalist needs an empirical interpretation for his abstract theory of functional systems; in other words, he needs to be able to identify at least some mechanisms within the physical system as realizations of abstract internal states.

Although it may be argued that reductionism is justified only after its program has been completed, this way of putting the problem is certainly too stringent, as it obviously amounts to dismissing all reductionist projects out of hand. It is clear, therefore, that the functionalist (and the identity theorist alike) need not offer a complete reduction scheme to warrant his claim; it will suffice if he presents a compelling sketch of it, and succeeds in showing that there are no grounds to doubt the possibility of pursuing his project to its end. It is crucial in this connection to recognize the difference between arguing for a particular reduction in science and attempting to show that some sort of reduction in a given domain is possible and fruitful.

Once the requisite theories are in place, the scientist may strive to attain a maximally informative reduction in the sense of committing himself to a number of particular theories in

[5] The disjunction in (2) is not exclusive.

order to provide detailed answers to various questions; such a reduction will be criticized if it is found to provide answers that are not detailed enough or if the details do not square with current scientific knowledge. In contrast, the best strategy to adopt when arguing for a reductionist research program is to make it as general and open-ended as possible, thereby accommodating for the various unforeseeable discoveries which are bound to appear in the course of future investigation. Here, any positive claim is a liability, prone to come into conflict with future theories and discoveries; thus, in discussing a reductionist research program, one must bear in mind that all such detailed claims are tentative and serve as illustrations rather than assertions. To be sure, a certain number of such tentative claims is indispensable if this kind of argument is to be at all compelling, but, again, the standards they must meet are not as high as most critics would like them to be.

In keeping with what has been said above, Putnam's proposed reduction is more general than the models he offered in his earlier papers. First of all, he introduces the notion of a Probabilistic Automaton, which is an extension of a Finite Automaton in that the Machine Table of a Probabilistic Automaton additionally specifies transition probabilities of passing from (a) input to state, (b) state to state, and (c) state to output. (A Deterministic Finite Automaton is, of course, a species of a Probabilistic Automaton, namely such that all its transition probabilities assume the value of 1.) This Probabilistic Automaton is equipped with sense and motor organs whose physical description is provided, though their states and inputs are specified only implicitly by the transition probabilities (these organs may be construed as physical realizations of the input-output tape).

Since any physical system can be described as a realization of many different Turing machines, the notion of a Description is needed. Putnam (1967b: 434) writes, 'A Description of S, where S is a system, is any true statement to the effect the S possesses distinct states S_1, S_2, … S_n which are related to one another and to the motor outputs and sensory inputs by the transition probabilities given in such-and-such a Machine Table.' The Machine Table referred to in the Description is called 'the Functional Organization of S relative to that Description,' while by 'the Total State of S (at a given time) relative to the Description' Putnam means 'the S_i such that S is in state S_i at that time.'

The thesis of strong functionalism is then explicated as follows (1967b: 434):

(1) All organisms capable of feeling pain are Probabilistic Automata.

(2) Every organism capable of feeling pain possesses at least one Description of a certain kind (i.e. being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).

(3) No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in (2).

(4) For every Description of the kind referred to in (2), there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in that subset.

This formulation is accompanied by the following explanations: condition (1) is redundant because any organism possesses a Description, condition (3) is there to exclude 'swarms of bees', conditions (2) and (4) must be made precise by future developments in science (we can now offer only tentative descriptions of what it is to be feeling pain, and what are its causes; condition (2) requires finding the right level of description, that is, passing from species-specific to species-independent models).

Before moving on I must remark that condition (3) can be dropped, as it is *ad hoc*. (To see this, it is enough to imagine that we were composed of small homunculi; it should be obvious that this would not automatically transport us outside the domain of psychology – see Block 1981b for details.) There is, also, a standard line of argument against strong functionalism which I am going to disregard. The kind of criticism I have in mind consists in protesting that this or that aspect of the mind cannot be reduced to Functional Organization; the reason I will not deal with it here is that answering such criticisms requires providing detailed analyses of a number of mentalistic terms; yet, such analyses are open to strictly empirical charges. In short, this would embroil us in a long and, ultimately, fruitless argument. Once again, we should remember that strong functionalism does not endorse a particular reduction – it merely argues for the adoption of a reductionist research program.

In keeping with what has been said earlier, the first task the proponent of strong functionalism faces consists in showing how his research program fares vis-à-vis identity theory.

One observation Putnam made on the basis of the multiple realizability argument was that it was difficult to suppose that there exist species-independent structural properties of the brain which can be identified with mental states. If he was right, then the same problem emerges for strong functionalism, which is, willy-nilly, committed to the possibility of isolating biological realizations of mental states. Suppose, for instance, that one claimed to be able to show that a cat and a man were in the same functional state when apparently experiencing a certain kind of pain. Since we know that, in both cases, the pain must be situated in the brain, then it follows that our strong functionalist would have to offer a description which revealed what the two brains had in common, which would be tantamount to asserting the existence of a structural species-independent property of the brain.

One way of avoiding this charge would be to insist that providing the Description of a man or cat does not require isolating the actual mechanisms which answer to it; presumably, if an organism seems to obey the laws of psychology then it does not matter how these laws are being realized. This reply will not do, however, unless we are willing to become behaviorists (it excludes super-Spartans, who do not exhibit pain-behavior).[6] The appeal of strong functionalism comes from the claim that the problem of other minds, as well as the mind-body problem, has an empirical solution, and so its adherent must be able to indicate what empirically discoverable mechanisms are responsible for realizing a given mental state.

It seems, then, that identity theorist and strong functionalist are in the same boat as regards the possibility of isolating species-independent properties of the brain which can be identified with mental states. It may appear that the boat is not a very comfortable one, so both kinds of theorists must find a different means of transport, but, fortunately, the empirical strand of the multiple realizability argument is flawed.

As Bechtel, Mundale (1999) point out, the problem is that the notion of brain state or brain structure operative in the multiple realizability argument is a philosopher's fiction. The concept which is closest to the reductionist's intuition can be identified with what neuroscientists call 'activity in the same brain part or conglomerate of parts'. That notion, however, is not as fine-grained as philosophers of mind often suppose it to be: brain maps constructed by neuroscientists are based not only on visible anatomical differences, but primarily on cytoarchitectonic and functional ones. Moreover, isolating brain regions in terms of function involves presupposing a psychological theory. In neuroscience, as it is practiced today, anatomical differences between conspecifics are discounted (that is, by the neurologist's lights, two people can definitely be in the same brain state, even if a PET scan shows that neural activity occurring in the first one's brain is of a different shape and in an apparently different region than that occurring in the brain of the second); also, brain regions are identified across various species, which makes it possible for the biologist to cite findings concerning cats or marmosets in support of a claim regarding humans.

The upshot is that we are, after all, entitled to claim that there are species-independent properties of the brain which can be identified with mental states.

Thus, we see considerable overlap between identity theory and strong functionalism; in fact, the only marked distinction between the two doctrines is that the former cannot cope with the possibility of brainless Martians.

As we remember, what I called the conceptual component of the multiple realizability argument is formulated as a thought-experiment whose conclusion is that, even if we knew that a

[6] It should be obvious that strong functionalism is compatible with behaviorism.

Martian did not have a carbon-based brain, we would nonetheless ascribe mental states to him. Functionalism is taken to explain such ascriptions by appeal to the principle that what we mean by 'mental states' can be defined functionally, or, more metaphysically speaking, that the essence of a mental state is relational.[7]

There are two crucial features of this argument which need emphasizing. First, it is an inference to the best explanation. What is being explained is multiple realizability itself; the tentative *explanans* is the doctrine of functionalism. The argument will succeed once it has been justified that functionalism provides a better account of the phenomenon of multiple realizability than other metaphysical positions. Second, being a thought-experiment, the conceptual component cannot tell us anything about the property of being a mental state – what it can reveal is that our intuitive notion of mental state is such that it permits multiple realizability. To put the same point differently, unless one takes some deep principles of folk psychology as true, the conceptual strand of the multiple realizability argument will only uncover some pre-theoretic intuitions about the mental that people fall back on when they lack sufficient information to make their decision on more solid grounds.

Having clarified what sort of conclusion we can expect from the argument, we may enquire whether functionalism is the best account of our intuitions. As much as I would like the answer to be in the positive, it looks as though the correct reply is 'no'. There is a number of thought-experiments that clash with functionalism. We may, for example, imagine a very simple organism having experiences or entertaining complicated thoughts – a functionalist must claim the fantasy to be somehow incoherent; such a situation should not be conceivable if we do operate with the functionalist conception of the mental. There is a complementary shortcoming: we have no difficulty conceiving a physical copy of ourselves which is deprived of mentality. Again, if functionalism reveals our deepest conceptual intuitions, we should not be able to do that. David Lewis's case of mad pain, a state phenomenologically similar to normal pain but caused by moderate exercise on an empty stomach and eliciting thoughts of mathematics, is no less puzzling (see Lewis 1981). To accommodate such examples the functionalist must engage in logical analyses of the kind proposed by logical behaviorists, analyses which Putnam had rightly rejected.

It seems much more plausible to suppose that our intuitive judgments concerning the ascription of mental states are driven by dualist intuitions. No matter how committed we might be to identity theory or functionalism, there remains, buried deep in our souls, a vestige of dualism which interferes with our philosophical beliefs, generating tensions and heated debates. The critique of materialism in Kripke 1971 is a good contemporary example of this; the

[7] My exposition of the multiple realizability argument owes a great deal to Ramsey 2006.

discussion of the problem of other minds in Mates 1981, where essentially the same point is made, is yet another.

Recognizing that dualism might be the best explanation of multiple realizability intuitions does not amount to its acceptance as a metaphysical doctrine; nor does it commit us to abandoning broadly construed materialism. In fact, if we take folk psychology to be an empirical theory, there are good reasons for rejecting many of its tenets together with its dualist underpinnings. What we should bear in mind is that identity theory and functionalism are not concerned with what our conceptual scheme is like; rather, they are claims about the structure of the world. And the world is the way it is regardless of how we picture it.

Returning to our brainless Martian, we see that the failure of identity theory to treat him in a way compatible with our intuitions is not as serious a shortcoming as some functionalists tend to portray it; the reason is that the physicalism–functionalism debate, correctly understood, lies in the domain of futurology rather than conceptual analysis. As both doctrines presuppose some future psychological theory which will eventually displace folk psychology, their disagreement springs from a difference of opinion regarding the future theory's conceptual apparatus. The functionalist claims that the future theory must retain multiple realizability, whereas the identity theorist asserts that it will not. There are interesting arguments on both sides, to be sure, but they are hardly conclusive.

One last aspect of reductionism is worth noting in this connection. So far, we have tacitly assumed, along with Putnam, that the future reduced theory will be as unified as folk psychology appears to be, even though it is at least logically possible that a number of theories will emerge, each accounting for a different set of mental phenomena. When applied to strong functionalism, this observation suggests a distinction between what I shall call *theory-functionalism* and *concept-functionalism*.

The former doctrine presupposes the existence of a unified, well-confirmed theory of the mental which is general enough to explain the behavior of earthly animals (including humans), robots and creatures from outer space. The theory-functionalist will claim that once such a theory is in place, it will be possible to show which concrete mechanisms realize which internal states identified by the theory, regardless of the physical nature of the system in question.

The latter position, by contrast, consists in a weaker claim; namely, that once a suitable theory of a given psychological concept (or, more precisely, a group of concepts) has been completed, it will be possible to show which concrete mechanisms in a given system instantiate the concept. In other words, supposing we have analyzed the concept of pain, we are entitled to seek its various physical instantiations, even though the analysis we employ may not be part of a larger psychological theory. It is well to note that concept-functionalism does not presuppose the existence of a unified set of psychological laws, nor is it committed to analyzing all psychological concepts on one level – it allows for the possibility that some mental states or processes are more biological (pain, hunger, fear, sexual desire), while others are more abstract (depression, envy, love, rational reflection); moreover, the laws governing mental states on one level may have little to do with laws holding of states on other levels.

The distinction between theory-functionalism and concept-functionalism can be easily mapped onto identity theory. Both concept-reductionisms are less adventurous than their theory-counterparts. More importantly, they permit philosophers to finesse their positions by blending functionalism and identity theory, or simply by restricting their reductionist claims to a particular group of concepts (see, for example, Fodor 2000). Lastly, Putnam's weak functionalism can be reconstructed as a species of concept-functionalism allowing for the correct theory of consciousness to be distinct from other psychological theories.

REFERENCES

Bechtel, William; Mundale, Jennifer
1999 'Multiple Realizability Revisited: Linking Cognitive and Neural States', *Philosophy of Science* 66, 175–207.

Block, Ned J. (ed.)
1981a *Readings in the Philosophy of Psychology*, vol. 1, The MIT Press, Cambridge, Ma.

Block, Ned J.
1981b 'Troubles with Functionalism', in Block (ed.), 269–305.

Block, Ned J.; Fodor, Jerry A.
1972 'What Psychological States are Not', *The Philosophical Review* 81, 159–81.

Fodor, Jerry A.
2000 *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, The MIT Press, Cambridge, Ma. – London.

Guttenplan, Samuel (ed.)
1994 *A Companion to the Philosophy of Mind*, Basil Blackwell, Oxford.

Kleene, Stephen Cole
1952 *Introduction to Metamathematics*, North-Holland Publishing Co., Amsterdam – P. Noordhoff N.V. – Groningen.

Kripke, Saul
1971 *Naming and Necessity*, Harvard University Press, Cambridge, Ma.

Lewis, David
1981 'Mad Pain and Martian Pain', in Block (ed.), 216–22.

Mates, Benson

1981 *Skeptical Essays*, University of Chicago Press, Chicago.

Quine, Willard van Orman

1950 'Two Dogmas of Empiricism', *The Philosophical Review* 60, 20–43.

Putnam, Hilary

1957 'Psychological Concepts, Explication and Ordinary Language', *The Journal of Philosophy*, 54, 94–9.

1960 'Minds and Machines', in Hook, S. (ed.), *Dimensions of Mind*, New York University Press, New York. Reprinted in Putnam 1975b, 362–85.

1963 'Brains and Behavior', in Butler, R. (ed.), *Analytical Philosophy Second Series*, Basil Blackwell, Oxford. Reprinted in Putnam 1975b, 325–41.

1964 'Robots: Machines or Artificially Created Life?', *The Journal of Philosophy* 61, reprinted in Putnam 1975b, 386–407.

1967a 'The Mental Life of Some Machines', in Castaneda, H. (ed.), *Intentionality, Minds and Perception*, Wayne State University Press, Detroit. Reprinted in Putnam 1975b, 408–28.

1967b 'The Nature of Mental States', first published as 'Psychological Predicates' in Capitan, W. H.; Merrill, D. D. (eds.), *Art, Mind and Religion*, University of Pittsburgh Press, Pittsburgh. Reprinted in Putnam 1975b, 429–40.

1969 'Logical Positivism and the Philosophy of Mind', in Achinstein, P.; Barker, S. (eds.), *The Legacy of Logical Positivism*, Johns Hopkins University Press, Baltimore. Reprinted in Putnam 1975b, 441–51.

1973c 'Reductionism and the Nature of Psychology', *Cognition* 2, reprinted in abbreviated version in Putnam 1994, 428–40.

1975a *Mathematics, Matter and Method. Philosophical Papers*, vol. 1, Cambridge University Press, Cambridge – London – New York – Melbourne.

1975b *Mind, Language and Reality. Philosophical Papers*, vol. 2, Cambridge University Press, Cambridge – London – New York – Melbourne.

1975d 'Other Minds', in Putnam 1975b, 342–61.

Ramsey, William

2006 'Multiple Realizability Intuitions and the Functionalist Conception of the Mind', *Metaphilosophy* 37, 53–73.

Ziff, Paul

1960 *Semantic Analysis*, Cornell University Press, Ithaca, NY.