The Rise and Fall of Computational Functionalism

Oron Shagrir

1. Introduction

Hilary Putnam is the father of computational functionalism, a doctrine he developed in a series of papers beginning with "Minds and machines" (1960) and culminating in "The nature of mental states" (1967b). Enormously influential ever since, it became the received view of the nature of mental states. In recent years, however, there has been growing dissatisfaction with computational functionalism. Putnam himself, having advanced powerful arguments against the very doctrine he had previously championed, is largely responsible for its demise. Today, Putnam has little patience for either computational functionalism or its underlying philosophical agenda. Echoing despair of naturalism, Putnam dismisses computational functionalism as a utopian enterprise.

My aim in this article is to present both Putnam's arguments for computational functionalism, and his later critique of the position.[1] In section 2, I examine the rise of computational functionalism. In section 3, I offer an account of its demise, arguing that it can be attributed to recognition of the gap between the computational-functional aspects of mentality, and its intentional character. This recognition can be traced to two of Putnam's results: the familiar Twin-Earth argument, and the less familiar theorem that every ordinary physical system implements every finite automaton. I close with implications for cognitive science.

2. The rise of computational functionalism

Computational functionalism is the view that mental states and events – pains, beliefs, desires, thoughts and so forth – are computational states of the brain, and so are defined in terms of "computational parameters plus relations to biologically characterized inputs and outputs" (1988: 7). The nature of the mind is independent of the physical making of the brain: "we could be made of Swiss cheese and it wouldn't matter" (1975b: 291).[2] What matters is our functional organization: the way in which mental states are causally related to each other, to sensory inputs, and to motor outputs. Stones, trees, carburetors and kidneys do not have minds, not because they are not made out of the right material, but because they do not have the right kind of functional organization. Their functional organization does not appear to be sufficiently complex to render them minds. Yet there could be other thinking creatures, perhaps even made of Swiss cheese, with the appropriate functional organization.

The theory of computational functionalism was an immediate success, though several key elements of it were not worked out until much later. For one thing, computational functionalism presented an attractive alternative to the two dominant theories of the time: classical materialism and behaviorism. Classical materialism – the hypothesis that mental states are brain states – was revived in the 1950s by Place (1956), Smart (1959) and Feigl (1958). Behaviorism – the hypothesis that mental states are behavior-dispositions – was advanced, in different forms, by Carnap (1932/33), Hempel (1949) and Ryle (1949), and was inspired by the dominance of the behaviorist approach

in psychology at the time. Both doctrines, however, were plagued by difficulties that did not, or so it seemed, beset computational functionalism. Indeed, Putnam's main argument for functionalism is that it is a more reasonable hypothesis than classical materialism and behaviorism.

The rise of computational functionalism can be also explained by the "cognitive revolution" of the mid-1950s. Noam Chomsky's devastating review of Skinner's *Verbal Behavior*, and the development of experimental instruments in psychological research, led to the replacement of the behaviorist approach in psychology by the cognitivist. In addition, Chomsky's novel mentalistic theory of language (Chomsky 1957), which revolutionized the field of linguistics, and the emerging research in the area of artificial intelligence, together produced a new science of the mind, now known as cognitive science. The working hypothesis in this science has been that the mechanisms underlying our cognitive capacities are species of information processing, namely, computations that operate on mental representations. Computational functionalism was inspired by these dramatic developments. Putnam, and even more so Jerry Fodor (1968, 1975), thought of mental states in terms of the computational theories of cognitive science. Many even see computational functionalism as furnishing the requisite conceptual foundations for cognitive science. Given its close relationship with the new science of the mental, it is not surprising computational functionalism was so eagerly embraced.

Putnam develops computational functionalism in two phases. In the earlier papers, Putnam (1960, 1964) does not put forward a theory about the nature of mental states. Rather, he uses an analogy between minds and machines to show that "the various issues

and puzzles that make up the traditional mind-body problem are wholly linguistic and logical in character… all the issues arise in connection with any computing system capable of answering questions about its own structure" (1960: 362). Only in 1967 does Putnam make the additional move of identifying mental states with functional states, suggesting that "to know for certain that a human being has a particular belief, or preference, or whatever, involves knowing something about the functional organization of the human being" (1967a: 424). In "The nature of mental states", Putnam explicitly proposes "the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism" (1967b: 433).

2.1 ***The analogy between minds and machines***

Putnam advances the analogy between minds and machines because he thinks that the case of machines and robots "will carry with it clarity with respect to the 'central area' of talk about feelings, thoughts, consciousness, life, etc." (1964: 387). According to Putnam, this does not mean that the issues associated with the mind-body problem arise for machines. At this stage Putnam does not propose a theory of the mind. His claim is just that it is possible to clarify issues pertaining to the mind in terms of a machine analogue, "and that all of the question of 'mind-body identity' can be mirrored in terms of the analogue" (1960: 362). The type of machine used for the analogy is the Turing machine, still the paradigm example of a computing machine.

A Turing machine is an abstract device consisting of a finite program, a read-write head, and a memory tape (figure 1). The memory tape is finite, though indefinitely extendable, and divided into cells, each of which contains exactly one (token) symbol from a finite alphabet (an empty cell is represented by the symbol B). The tape's initial configuration is described as the 'input'; the final configuration as the 'output'. The read-write mechanism is always located above one of the cells. It can scan the symbol printed in the cell, erase it, or replace it with another. The program consists of a finite number of states, e.g., A, B, C, D, in figure 1. It can be presented as a machine table, quadruples, or, as in our case, a flow chart.

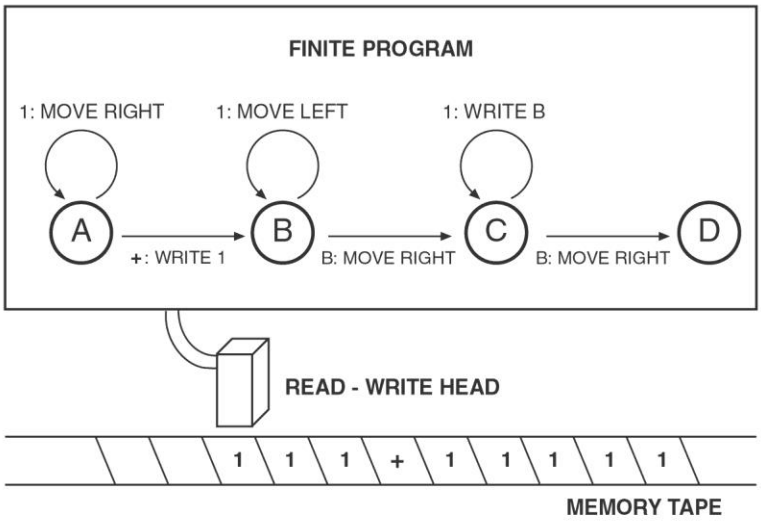

Figure 1: A Turing-machine

The computation, which mediates an input and an output, proceeds stepwise. At each step, the read-write mechanism scans the symbol from the cell above which it is located, and the machine then performs one or more of the following simple operations: (1) erasing the scanned symbol, replacing it with another symbol, or moving the read-

write mechanism to the cell immediately to the right or left of the cell just scanned; (2) changing the state of the machine program; (3) halting. The operations the machine performs at each step are uniquely determined by the scanned symbols and the program's instructions. If, in our example, the scanned symbol is '1' and the machine is in state A, then it will follow the instruction specified for state A, e.g., 1: R, meaning that it will move the read-write mechanism to the cell immediately to the right, and will stay in state A.

Overall, any Turing machine is completely described by a flow chart. The machine described by the flow chart in figure 1 is intended to compute the function of addition, e.g., '111+11', where the numbers are represented in unary notation. The machine starts in state A, with the read-write mechanism above the leftmost '1' of the output. The machine scans the first '1' and then proceeds to arrive at the sum by replacing the '+' symbol by '1', and erasing the rightmost '1' of the input. Thus if the input is '111+11', the printed output is '11111'.

The notion of a Turing machine immediately calls into question some of the classic arguments for the superiority of minds over machines. Take for example Descartes' claim that no machine, even one whose parts are identical to those of human body, cannot produce the variety of human behavior: "even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others" (1637/1985: 140). It is true that our Turing machine is only capable of computing addition. But as Turing proved in 1936, there is also a universal Turing machine capable of computing any function that can be computed by a Turing machine.

In fact, almost all the computing machines used today are such universal machines. Assuming that human behavior is governed by some finite rule, it is hard to see why a machine cannot manifest the same behavior.[3]

As Putnam shows, however, minds and Turing machines are not just analogous in the behavior they are capable of generating, but also in their internal composition. Take our Turing machine. One characterization of it is given in terms of the program it runs, i.e., the flow chart, which determines the order in which the states succeed each other, and what symbols are printed when. Putnam refers to these states as the "logical states" of the machine, states that are described in logical or formal terms, not physical terms (1960: 371). But "as soon as a Turing machine is physically realized" (ibid.) the machine, as a physical object, can also be characterized in physical terms referring to its physical states, e.g., the electronic components. Today, we call these logical states 'software' and the physical states that realize them 'hardware'. We say that we can describe the internal makeup of a machine and its behavior both in terms of the software it runs (e.g., WORD), and in terms of the physical hardware that realizes the software.

Just as there are two possible descriptions of a Turing machine, there are two possible descriptions of a human being. There is a description that refers to its physical and chemical structure; this corresponds to the description that refers to the computing machine's hardware. But "it would also be possible to seek a more abstract description of human mental processes in terms of 'mental states'… a description which would specify the laws controlling the order in which the states succeeded one another" (1960: 373). This description would be analogous to the machine's software: the flow chart that

specifies laws governing the succession of the machine's logical states. The mental and logical descriptions are not similar only in differing from physical descriptions. They are also similar in that both thought and 'program' are "open to rational criticism" (1960: 373). We could even design a Turing machine that behaves according to rational preference functions (i.e., rules of inductive logic and economics theory), which, arguably, are the very rules that govern the psychology of human beings; such a Turing machine could be seen as a rational agent (1967a: 409-410).

There is thus a striking analogy between humans and machines. The internal makeup and behavior of both can be described, on the one hand, in terms of physical states governed by physical laws, and on the other, more abstractly, in terms of logical states (machines) or mental states (humans) governed by laws of reasoning. Putnam contends that this analogy should help us clarify the notion of a mental state, arguing that we can avoid a variety of mistakes and obscurities if we discuss questions about the mental – the nature of mental states, the mind-body problem and the problem of other minds – in the context of their machine analogue. Take, for example, the claim that if I observe an after-image, and at the same time observe that some of my neurons are activated, I observe two things, not one. This claim supposedly shows that my after-image cannot be a property of the brain, i.e., a certain neural activity. But, Putnam (1960: 374) observes, this claim is clearly mistaken. We can have a clever Turing machine that can print 'I am in state A', and at the same time (if equipped with the appropriate instrumentation) print 'flip-flop 36 is on' (the realizing state). This, however, does not show that two different events are taking place in a machine. One who nonetheless draws

the conclusion from the after-image argument that souls exist, "will have to be prepared to hug the souls of Turing machines to his philosophical bosom!" (1960: 376).

2.2 The functional nature of mental states

In 1967a and 1967b, Putnam takes the analogy between minds and machines a step further, arguing that pain, or any other mental state, is neither a brain state nor a behavior-disposition, but a functional state. Before looking at the notion of a functional state (section 2.2.2) and at Putnam's specific arguments for functionalism (section 2.2.3), let us elucidate the context in which these claims are made.

2.2.1 ***Is pain a brain state?***

In 1967b, Putnam raises the question: what is pain? In particular, is it a brain state? On the face of it, the question seems odd. After all, it is quite obvious, even if hard to define, what pain is. Pain is a kind of subjective conscious experience associated with certain 'feel' ('quale' in philosophical parlance). Even Putnam agrees that pain is associated with a certain unpleasant conscious experience: "must an organism have a *brain* to feel pain?" (1967b: 439). Why, then, does Putnam question what pain is, and what could be his motivation for wondering if pain could be something else, e.g., a brain state?

To inquire into the definition of pain is to try and identify that which is common to all pains, or that which is such as to render a certain phenomenon pain. At a more

general level, philosophers seek the ultimate mark of the mental: the feature that distinguishes mental from non-mental phenomena. Conscious experience is often deemed that which is characteristic of the mental. Other serious contenders are intentionality (Brentano), rationality (Aristotle), and disposition (Ryle). And even if no single such mark exists, it is nonetheless edifying to explore the relations between the different aspects of mentality.

Functionalism is, roughly, the view that the mark of the mental has to do with the role it plays in the life of the organism. To help us grasp the functionalist account of the mental, it may be useful to consider functionalist definitions of other entities. A carburetor is an object defined by its role in the functioning of an engine (namely, mixing fuel and air). A heart is defined by the role it plays in the human body (namely, pumping blood). The role each object plays is understood in the context of the larger organ it is part of, and is explicated in terms of its relations to the other parts of that organ. The material from which the object is made is of little significance, provided it allows the object to function properly. Similarly, the functionalist argues, mental states are defined by their causal relations to *other mental states*, sensory inputs and motor outputs. An early version of functionalism is sometimes attributed to Aristotle. Some versions of functionalism are popular in contemporary philosophical thinking. Computational functionalism is distinguished from other versions of functionalism in that it explicates the pertinent causal relations in terms of computational parameters.[4]

Some philosophers require that the distinguishing mark of pain be described in ‘non-mental’ terms, e.g., physically, neurologically, behaviorally or even formally. These

philosophers ask what pain is, not because they deny that pain is associated with a subjective conscious experience, but because they maintain that if pain is a real phenomenon, it must really be something else, e.g., C-fiber stimulation. The task of the philosopher, they argue, is to uncover the hidden nature of pain, which, they all agree, is indeed, among other things, an unpleasant conscious experience. Such accounts of mental states are called naturalistic or *reductive*. While Aristotle's version of functionalism is not reductive, computational functionalism has always been conceived as a reductive account. Indeed, in advancing computational functionalism, Putnam sought to provide a reductive alternative to the reigning reductive hypotheses of the time: classical materialism and behaviorism.

Having considered why a philosopher would ask whether pain is a brain state, let us now consider what would constitute an admissible answer: under what conditions would we affirm that pain *is* a brain state (or a behavior disposition, or a functional state)? It is customary in contemporary philosophy of mind to distinguish two senses of the claim that 'pain is a brain state', one at the level of events (token-identity), another at the level of properties (type-identity). At the level of events, 'pain is a brain state' means that any token of pain – any event that is painful – is also a token of some brain activity. At the level of properties, 'pain is a brain-state' means that the property of being painful is identical with some property of the brain, e.g., C-fiber stimulation. Token-identity does not entail type-identity. It might be the case that *any* pain token is *some* brain-state in the sense that it has neurological properties, though there is no single neurological property that applies to all pain tokens. My pain could be realized in C-fiber stimulation, whereas

that of other organisms is realized in very different brain states. It is important to see that Putnam's question about pain and brain-states is framed at the level of *properties*, not events. The question Putnam is asking is whether the property of being in pain is identical with some property of the brain.[5]

We still have to say something about identity of properties. On what basis would we affirm or deny that pain is a property of the brain (or a type of behavior-disposition or a functional property)? Putnam is undecided on the issue in his earlier papers (1960, 1964, 1967a), but in 1967b settles on the view that the truth of identity claims such as 'pain is C-fiber stimulation' is to be understood in the context of *theoretical identification*. The inspiration comes from true identity claims such as 'water *is* H_2O', 'light *is* electromagnetic radiation' and 'temperature *is* mean molecular kinetic energy'. In saying that 'water is H_2O', we assert that: (a) The properties of being water and being H_2O molecules are the same in the sense that they apply to exactly the same objects and events. Or at the linguistic level, that the terms 'water' and 'H_2O' (which 'express' the properties) are coextensive. (b) The terms have the same extension (or the properties apply to the same objects/events) not only in our world, but in every possible physical world. They are, roughly speaking, necessarily coextensive. Their coextensiveness is a matter of the laws of science. (c) Affirming that they are coextensive is likely to be a matter, not of conceptual analysis (one could think about water yet know nothing about molecules of H_2O), but of empirical-theoretical inquiry. The inquiry is empirical in the sense that it was *discovered*, by way of scientific research, that the extension of 'water', namely, the stuff that fills our lakes, runs in our faucets, etc., is H_2O. And it is theoretical

in the sense that familiar explanatory practices enjoin us to deem the empirical coextensiveness identity.

Similarly, to say that 'pain is C-fiber stimulation' is to assert that the two properties apply to the same class of physically possible events and states. Or in other words, that the terms 'pain' and 'C-fiber stimulation' refer, necessarily, to exactly the same states and events. Yet this assertion is not likely to be determined by a conceptual analysis (one could think about pain yet know nothing about C-fiber stimulation). 'Pain is C-fiber stimulation' is a hypothesis whose truth-value is likely to be ascertained through empirical-theoretical research, possibly conducted by cognitive scientists.

In sum, then, pain *is* a brain state (or a behavior-disposition, etc.) just in case there is a brain-property (or a kind of behavior-disposition, etc.) Q, such that the following two conditions hold:

Unique realization (UR_Q): any physically possible pain-event is also a Q-event (event of type Q).

Supervenience (SUP_Q): any physically possible Q-event is also a pain-event.

The first condition, UR_Q, asserts that all pains, actual and possible, are realized in events of type Q. I call this condition Unique Realization to signify that there cannot be two organisms, both of which feel pain, but one of which has Q, and the other not. Pain is always realized in Q-events. The second condition, SUP_Q, complements the first. It asserts that all Q-events are realizations of pain. I call this condition Supervenience to signify that there cannot be two organisms, both of which have exactly the same property

Q, only one of which feels pain. Being in pain is determined by, that is, dependent on, having Q.

2.2.2 ***What is computational functionalism?***

Putnam sets out the concepts and ideas underlying computational functionalism as far back as 1960. In a discussion of computing machines, Putnam mentions that the term 'functional organization' is used to describe a computing machine in terms of sequences of logical states (1960: 373). He also mentions that logical states are characterized in terms of their "*relations* to each other and to what appears on the tape" (1960: 367). And he emphasizes that this characterization is expressed in logical-mathematical language, e.g., a flow chart description, that makes reference *neither* to the 'physical realization' of the logical states, in copper, platinum, etc. (1960: 367, 373), *nor* to the interpretation given to the symbols. For example, being in state B of our Turing machine is represented by the following 'maximal' description:

> Being in B: being in the second of four states S_1, S_2, S_3, S_4 that are related to one another and to inputs and outputs as follows. If being in S_1, then getting '1' as an input results in moving one cell to the right; and getting '+' as an input results in writing '1' as an output and going to S_2. If being in S_2, then getting '1' as an input results in moving one cell to the left; and getting B as an input results in moving one cell to the right, and going to state S_3. If being in S_3, then getting '1' as an

input results in writing B as an output; and getting B as an input results in moving one cell to the right. If being in S_4, then halting.[6]

In 1967, Putnam takes the additional step of identifying the mind with the functional organization of thinking organisms, and mental states with functional states: "being capable of feeling pain *is* possessing an appropriate kind of Functional Organization" (1967b: 434). This move encompasses two claims: *computationalism* and (computational) *functionalism* (henceforth, I will use 'functionalism' to denote computational functionalism). Computationalism is the claim that organisms with minds have functional organization, i.e., there is a true 'flow chart' description of the organism in terms of states and their relations to each other and to inputs and outputs. Functionalism is the claim that having a mind is having the right sort of functional organization, and any mental property is a certain *kind* of this functional organization. This means that being in pain is having some property that is characteristic of this functional organization. More generally, for any mental property M there is a functional property F such that the following two conditions hold:

UR$_F$: any M-event is also an F-event.

SUP$_F$: any F-event is also an M-event.

Given what we know about the functional organizations of machines, functionalism has two important consequences. One consequence is that pain, as a state of the functional organization, is defined by its causal relations to other states (e.g., the belief that I am in pain), inputs (e.g., being punched), and outputs (e.g., the vocalization

'ouch'). The other is that the specification of pain is *reductive* in the sense that it is formulated in non-mental terms. That is, the specification of a mental state in terms of other mental states is eliminated in favor of a formula that contains logical terms (e.g., 'there is', 'and'), variables (i.e., x, S_1,…, S_n), and biological/physical terms (for the inputs and outputs), but no mental terms.

To see how the elimination works, assume that FO(S_1,…, S_n, i_1, …, i_k, o_1, …o_l) is my functional organization, namely, a full description of the relations between my internal states S_1, …, S_n, sensory inputs i_1, …, i_k, and motor outputs o_1, …o_l. The functionalist claim is that my being in pain is a state, say S_5, of this functional organization, and that any other organism is in pain just in case this organism has this (or an isomorphic) FO, and is in S_5. Thus being in pain can be specified as follows:

> Being in pain = being the fifth of n states, S_1,…, S_n, whose relations to one another and to inputs and outputs are specified by FO(S_1,…, S_n, i_1, …, i_k, o_1, …o_l).

Thinking about next summer's vacation is defined by the same formula, except that 'being in state S_5' is replaced with 'being in state S_{87}', and so forth.

While the characterization of mental states is analogous in some respects to the characterization of the logical states of a Turing machine, there are also important respects in which the characterizations differ. One such respect is the mode of specification of inputs and outputs. The inputs and outputs of Turing machines are

specified in syntactic terms (e.g., ‘1’). But this specification is much be too liberal to be used for the purposes of characterizing mental states, both because it does not fix the semantics of mental states (see below), and because it may also be true of other complex organizations, such as the economics of Liberia, that lack any mentality whatsoever (Block 1979). To remedy this situation, Putnam is careful to specify sensory inputs and motor outputs in physical or biological terms.

In “Philosophy and our mental life”, Putnam (1975b) also modifies his earlier claim that mental states are, literally, states of a complex Turing machine. One reason for the modification is that a Turing machine model cannot perspicuously represent learning, memory (1975b: 298-299; see also 1992a: 8-14 and 1997: 34). Another is that when one is in a state of pain, one is also in many other mental states (e.g., the state of believing that one is in pain), but a Turing machine instantiates only a single state at any given time (see also Block and Fodor, 1972). These modifications do not undermine functionalism. Functionalism is committed to the claim that mental states are computational states, not to the Turing machine model. It might well be that the functional organization of cognizing organisms is best represented in terms of neural networks, and not in terms of Turing machines (see, e.g., Churchland and Sejnowski 1992).

Computationalism is often associated with the maxim that the brain is a sort of computer, and as such, runs a program (“software”). Functionalism is commonly associated with the maxim that the mind *is* the software of the brain. But why should we believe either of these claims? Putnam does not provide a detailed argument for computationalism. He feels little need to do so, since he takes computationalism to be

"obviously, redundant, and only introduced for expository reasons… since everything is a Probabilistic Automaton under *some* description" (1967b: 435).[7] And even if it is not the case that everything can be seen as some kind of probabilistic automaton, cognitive science nonetheless insists that cognizing organisms are species of computing machines.[8] As for functionalism, the argument here is that it does a better job than the other reductive accounts, namely, classical materialism and behaviorism (1967a, 1967b). In fact, functionalism can be seen as correcting the deficiencies of classical materialism and behaviorism. Let us see why.

2.2.3 ***Functional states, brain states, and behavioral dispositions***

Classical materialism, recall, is the claim that any mental property is a property of the brain. Take pain. Pain is a brain property P (e.g., C-fiber stimulation) just in case the following conditions are met: (a) UR_P: any organism that is in pain has P, and (b) SUP_P: any organism that has P is in pain. Putnam challenges UR_P on the grounds that there may be an organism that feels pain, but in which pain is realized very differently than it is in humans. If, say, human pain is realized in C-fiber stimulation, but the other organism has no C-fibers at all. And after all, it is very likely that the brains of mammals, reptiles and mollusks are in very different physical-chemical states when these organisms are in pain (1967b: 436). All this is still consistent with the materialist's claim that any pain token is also a physical-chemical event. What is being denied is the further claim that the property of being in pain is a physical-chemical property. It is much more reasonable to assume

that different tokens of pain are realized in events with different physical-chemical properties.

Functionalism, on the other hand, is consistent with the multiple realizability of the mental. For what we learn from the case of machines is that "any Turing machine that can be physically realized at all can be realized in a host of totally different ways" (1967a: 418). In fact, if functionalism is true, and mental states are computational states, then it *must* be physically possible for them to be multiply- realizable: "our mental states, e.g., *thinking about next summer's vacation*, cannot be *identical* with any physical or chemical states. For it is clear from what we already know about computers etc., that whatever the program of the brain may be, it must be physically possible, though not necessarily feasible, to produce something with the same program but quite a different physical and chemical constitution" (1975b: 293). So with respect to the multiple realization of mental properties, functionalism is much further ahead than classical materialism. It is consistent with the materialist's claim that any pain token is likewise a token of a physical state, but *also* consistent with the claim that being in pain is not a brain property.

Putnam elaborates his argument against behaviorism in his "Brains and behavior" (1963). On the behaviorist account, pain is a kind of behavioral disposition: the disposition to emit certain responses (e.g., 'ouch') under certain stimuli (e.g., being punched in the face). In a sense, behaviorism welcomes the idea that pain is defined by its functional role, defined, that is, by the responses the organism produces under certain stimuli. As Putnam argues, however, behavioral dispositions do not successfully

explicate the concept of pain. Pain is a kind of behavioral-disposition B just in case (a) UR_B: any organism that is in pain is disposed to behave in the B-way, e.g., to emit the sound 'ouch' when punched in the face, and (b) SUP_B: any organism that is disposed to behave in the B-way is in pain. But behaviorism fails on both counts. UR_B fails because there might be a community of super-Spartans who, though they feel pain, are trained never to utter the sound 'ouch' when punched in the face. SUP_B fails because there could well be perfect actors who can display the same behavioral dispositions we do when we are injured even if their pain-fibers have been surgically removed.

What these and other examples demonstrate is that my pain-behavior is not just a result of my being in pain, but also of my being in *other* mental states, e.g., that of believing that uttering the sound 'ouch' is not outrageous behavior. We can, indeed, address this deficiency of the behavioristic account by admitting that pain is not just the disposition to utter the sound 'ouch' when punched in the face, but the disposition to utter the sound 'ouch' when punched in the face and when in *other mental states.* But this correction is tantamount to endorsing functionalism: pain is not just identified by the relations between stimuli and responses, but by the relations between stimuli, responses and mental states. Functionalism does away with the aforementioned counterexamples to behaviorism easily. The super-Spartan's pain is related to other mental states, hence he or she may react in a manner unlike ordinary humans. And the 'pain behavior' of the perfect actor results from mental states other than pain. Compared to classical materialism and behaviorism, functionalism seems to win hands down.

3. Troubles with functionalism

Upon emerging as the received view of the mental, functionalism became the focus of increasing scholarly attention. Many philosophers advanced arguments *against* it. I will not survey all the arguments here; for an exhaustive survey, see Block (1996). My aim is to present Putnam's line of argument against functionalism, which, I believe, goes a long way toward explaining its demise. The argument consists of two steps. It is first argued that there is a gap between mental states' functional-computational properties, on the one hand, and their intentional aspects, on the other, meaning that either UR_F or SUP_F is false. For example, Putnam, (1988, 1992b) argues against UR_F, by pointing out that the same thought can be realized in different computational structures. The argument is simple: functionalism is a holistic theory on which a mental state is defined by its causal relations to other mental states. But it is quite possible that two individuals, John and Mary, though somewhat different in functional organization (for Mary believes some proposition John does not), both believe that water is wet. Thus either we cannot attribute to John and Mary, or any other pair of individuals, the same belief--which is patently absurd--or we must admit that the same belief can be realized in different functional states. But if the latter is the case, then UR_F appears to be false: the same mental property can indeed be realized in different functional organizations (1988: 80-84, 1992b: 448-453).

Below, I discuss in some detail two further arguments Putnam makes against SUP_F. Both seek to show that the functional-computational properties of a mental state do not fix its intentional charachter. One is the well-known Twin-Earth argument (3.1), the

other the realization problem (3.2) that has received more attention of late. The upshot of all three arguments is that the functionalists must revise their initial characterization of mental states. They have to find an equivalence relation among the different types of functional organization, something they all have in common. Here, however, the second step in the argumentation kicks in. The second step is to argue that there is no hope the functionalist can specify such an equivalence relation yet preserve key elements of the theory. In particular, there is no hope of specifying the equivalence relation in non-intentional terms, something that is essential if the reductive character of the theory is to be preserved.

For example, to avoid the above-mentioned argument against UR_F, the functionalists have no choice but to adopt a less fine-grained individuation scheme. There are two routes the functionalists can take. One is fix some of the beliefs, e.g., that water is wet, as analytic. This would make these beliefs' individuation involenerable to realizations in different functional organizations, and affect the individuation of the other beliefs as well. But this move is not only undesirable, but won't help anyway (1988: 81-82, 1992b: 450-451; see also Fodor and Lepore, 1992). The other is to appeal to physical facts, a move that would blur the differences between functionalism and classical materialism, and turn functionalism into a utopian enterprise (1988, chapters 5 and 6; 1992b). I expand on this line of argumentation below, pointing to the difficulties that would beset any attempt to rehabilitate functionalism (3.3).

3.1 ***Content, computation and Twin-Earth***

That there might be a gap between the functional and the intentional – that the former may not determine the latter – should have been clear from the outset. After all, it is obvious that any computer program can be interpreted in different ways. One user may construe the program as playing chess, and another, as calculating the next month's payroll. What reason is there, then, to think that the program our brains run definitively determines the content of our thoughts, beliefs and other intentional states?

Let us be clearer about this multiple interpretation problem. A computer program is a formal-mathematical description containing only logical and mathematical operations that are defined over a finite set of symbols (e.g., '1', '0'). The symbols do, however, also have a semantic dimension. We take the machine in figure 1 to compute addition because we interpret the numeral '1' as representing the number ONE, and '0' as representing the number ZERO. When talking semantics, there are two elements that have to be taken into account. One is the symbol's extension, which is just the object or set of objects to which the symbol refers. Thus the extension of the numeral '1' is the number ONE, and the extension of 'water' is the set of things that consist of H_20 molecules. The other dimension is the symbol's content, which is what makes the symbol into the representation it is. It is the content of the symbol '1' that makes it representing ONE and not ZERO. Content is often associated with 'meaning', 'sense' and 'intension', and its nature is highly disputable. It is much less disputable, however, that (a) two symbols, i.e., 'water' and 'H_2O', can differ in their contents, yet have the same extension, and (b) if

two symbols have the same content, they also have the same extension. In a nutshell, content determines extension, but not vice versa.

The problem is now evident: there is a tension between the claim that mental content is computational-functional and the claim that mental content determines extension. On the one hand, if functionalism is correct, then the mental is, in its entirety, functional. In particular, the content of our thoughts, beliefs, and so forth, is exhaustively specified by their computational-functional properties. These are the functional properties of my thought that water is wet, for instance, that determine that I'm thinking about water and not cats. In other words, if functionalism is correct, then the content of our thoughts must supervene on their functional properties. If two organisms have exactly the same functional organization, then the content of their thoughts, beliefs, desires, etc, must be the same. And a fortiori, their thoughts must be about the same things. If the two organisms say that water is wet, both are thinking about water and not about cats: the extension of their concept WATER is the set of molecules of H_2O, not the set of cats.

But on the other hand, we know from the case of machines that the program – as a formal-syntactic entity – does not determine the extension of the symbols over which the operations are defined. In our toy example, we could take the '1' to stand for ZERO and the '0' for ONE, in which case the function computed would be quite different. We could also take the '1' and '0' to represent kinds of animals and not numbers at all. Functional organization constrains the set of possible interpretations, but does not determine a unique interpretation. It is always possible for two machines to have the same functional organization though their users interpret the symbols over which their operations are

defined quite differently. Similarly, it appears, two thinking organisms can be alike in functional organization though the extensions of what they say and think differ. If so, then computational functionalism is false: The contents of our thoughts do not supervene on functional properties.

It might be suggested that a sufficiently complex functional organization has a single interpretation. I think about water and not cats because the complexity of the program my brain runs rules out any non-hydrous content. But this suggestion won't work. Assuming that the program is complex enough (and formulated in a first-order language), it is guaranteed (by the Löwenheim-Skolem results) that the organization will have several non-isomorphic interpretations (see, e.g., Putnam 1980). This would explain why Putnam insists on biologically specified inputs and outputs (I/O). The hope is that if the specification of I/O is biological/physical and not merely formal-syntactic, this will rule out non-standard interpretations. I think about water and not cats since the physical perceptual inputs associated with this thought fix the hydrous content and rule out the feline content. Functional organization, then, consists of an implemented program (abstract automation) plus physically specified I/O.

In "The meaning of 'meaning'", however, Putnam (1975c) advances an argument whose upshot is that the appeal to physical I/O will not help either. Two individuals, Oscar and Toscar, can have exactly the same functional organization, including the same physically specified sensory and motor I/O, yet their concepts, thoughts, beliefs, desires and so forth have different contents. To see this, imagine that Toscar lives on Twin-Earth, which is exactly like Earth, except that the term 'water' refers to a liquid with the

chemical structure XYZ, a liquid that is thus very different from H_2O. Oscar and Toscar, however, cannot tell XYZ from H_2O, as the two liquids look, taste, smell and sound exactly the same. On Twin-Earth they have XYZ in those places where we have H_2O: rivers, clouds, faucets, and so on. It is thus possible for Oscar and Toscar to have exactly the same functional organization though their thoughts differ considerably in content. When Oscar says 'water is wet', he is referring to the liquid that is H_2O, whereas Toscar, when saying 'water is wet', refers to the liquid that is XYZ. Let us assume that Oscar and Toscar know nothing about H_2O or XYZ, as they live prior to 1750, when no one knew the chemical structure of the liquids. What Oscar and Toscar know is that 'water' refers to a liquid that is familiar in their respective environments. Yet the liquids to which Oscar's and Toscar's thoughts refer are very different. But given that content determines extension, and that the extensions are different, Oscar's and Toscar's thoughts must differ in content. Hence, mental contents do not supervene on functional properties.

The Twin-Earth argument created a storm in the philosophical community, reviving the view known as psychological externalism, on which some determinants of mental content are located in the speaker's environment. Some functionalists have argued in response that the argument only shows that content comprises two factors (Indeed, Putnam [1975c] himself suggested something along these lines, though he has since repudiated that view [1992b]). One factor determines extension, and is associated with "meaning". This factor is "wide", in the sense that some of its identity conditions make an essential reference to the individual's environment. The other factor is associated with features having to do with psychological/phenomenal properties. This factor is "narrow",

in the sense that it is not wide. This factor can still be identified functionally. The two-factor account is no longer popular, perhaps because it has proved difficult to explain how the factors are related, or due to the convincing arguments that have been advanced for the thesis that theories in cognitive science utilize "wide" individuation (e.g., Burge, 1986).

Other functionalists have suggested that I/O should be understood as extending all the way to the distal environment; this view is known as wide or global functionalism (e.g., Harman, 1988). The thoughts of Oscar and Toscar differ in content because they are causally related to different I/O. Oscar's thoughts are related to water, Toscar's, to "twater". As it turns out, however, the troubles for functionalism do not stop here.

3.2 The realization problem

In the appendix to *Representation and Reality* (1988: 121-125), Putnam proves that every ordinary open system is a realization of every abstract finite automaton. This result clearly threatens SUP_F. A corollary of the theorem is that there can be two objects, a human and a rock, with the same functional organization (save the physical I/O), only one of which is deemed to have mentality. Differently put, if a functional organization of a certain complexity is sufficient for having a mind, as the functionalist claims, then the rock too should be deemed to have a mind. In fact, almost everything, given that it realizes this automaton, has a mind. Moreover, if Putnam's theorem is true, then my brain simultaneously implements infinitely many different functional organizations, each

constituting a different mind. It thus seems that I should simultaneously be endowed with all possible minds!

Putnam's theorem therefore seems to undermine SUP_F in two ways: (a) A rock implements the same functional organization I do, but the rock is deemed lacking in mentality. (b) My brain implements many functional organizations, each of which normally constitutes an independent mind, yet I have but a single mind. I will call these results the realization problem.

The realization problem cuts deeper than the Twin-Earth problem. Even if, adopting a less fine-grained scheme of individuation, we take the thoughts of Oscar and Toscar to be the same, this would do little to alleviate the realization problem. It would still be the case that the rock implements the functional organization of Oscar (and Toscar), and so should be deemed endowed with their minds. And it would still be the case that my brain implements all the functional organizations that constitute other minds. One reason the realization problem is so interesting is that it highlights the Computationalism thesis. Initially, Computationalism seemed innocuous, in that it seemed highly plausible that a cognizing organism would be functionally organized: that it would realize a probabilistic automaton of some sort. It turns out, however, that not only does such an organism have one functional organization, it has infinitely many: it realizes every finite automaton, and perhaps even many other kinds of automata, something that leads directly to the realization problem.

3.2.1 Outline of Putnam's proof

Putnam's theorem pertains to abstract finite state automata (FSA) without inputs/outputs. Take the FSA that runs through the state-sequence ABABABA in the time interval we want to simulate. Here A and B are the states of the FSA. Assume that a rock can realize this run in a 6-minute interval, say from 12:00 to 12:06. Assume that the rock is in a maximal physical state S_0 at 12:00, S_1 at 12:01, and so forth (a maximal physical state being its total physical makeup specified in complete detail). Also assume that the states differ from each other (this is Putnam's Principle of Noncyclical Behavior). Now let us define a physical state **a** as S_0 v S_2 v S_4 v S_6, and state **b** as S_1 v S_3 v S_5. The rock implements the FSA in the sense that the causal structure of the rock "mirrors" the formal structure of the FSA. The physical state **a** corresponds to the logical state A, the physical **b** corresponds to the logical B, and the causal transitions from **a** to **b** correspond to the computational transitions from A to B. A complete proof would require further elaboration, as well as a Principle of Physical Continuity. But the idea is wonderfully simple, and can be extended to any I/O-less FSA.

Putnam notes (1988: 124) that the proof cannot be immediately extended to FSA with I/O. If the I/O are functionally individuated, then the I/O can be treated much like abstract internal states, and the extension is more natural. But if the I/O are specific kinds of physical organs, as the functionalist requires, then the rock, which lacks motor or sensory organs of the required sort, cannot realize the automaton. The rock cannot implement a mind because it lacks the motor and sensory organs of thinking organisms.

The functionalist is in real trouble, I would argue, if the difference between rocks and humans is exhausted by the I/O issue. After all, the whole point of functionalism is that the difference between thinking organisms and rocks is rooted in the complexity of functional organization. But if Putnam's argument is correct, the difference between humans and rocks does not lie in the complexity of their respective internal computational arrangements (which turn out to be the same for humans and rocks), but in the kinds of I/O they can handle. Is this not behaviorism in a new guise?

Indeed, Putnam points out (1988: 124-125) that reliance on physical I/O causes functionalism to collapse into behaviorism. Functionalism improves on behaviorism by taking into account not only I/O, but also the mediating algorithm. Behaviorism is false because there are beings that have the same I/O dependencies as humans, but different mentalities, e.g., super-Spartans; or are altogether lacking in mentality (Block, 1995). Functionalism holds that the reason they differ is that the 'internal' relations among the logical states differ, that is, the implemented algorithm differs. But consider a physical object with the right kind of I/O, e.g., a super-Spartan. What is true of this entity, on Putnam's theorem, is that it realizes all the possible algorithms that mediate the I/O. There can be no difference between the algorithms implemented by this entity and the algorithms implemented by humans. "In short, 'functionalism', if it were correct, would imply behaviorism! If it is true that to possess given mental states is simply to possess a certain 'functional organization', then it is also true that to possess given mental states is simply to possess certain behavior dispositions!" (1988:124-125).

3.2.2 Chalmers's reply

Many attempted to downplay Putnam's result, arguing that it takes more than Putnam allows to implement the functional organizations that are minds. David Chalmers (1996) provides a detailed counterargument along these lines. His contention is that there are constraints on the notion of implementation that are not taken into account by Putnam, constraints not satisfied by rocks. For one thing, the state transitions of the implementing machine must be reliable and counterfactual supporting. For another, the causal structure of the physical object should mirror all the possible formal state transitions of the implemented FSA. In Putnam's proof, the rock implements only a *single run* (the transition from A and B and back), but not other runs that might exist. If the FSA has other state transitions, e.g., C→D and D→C, these transitions should also be mirrored by the rock's dynamics.

It thus follows, according to Chalmers, that Putnam's proof applies to relatively simple kinds of automata, but not to the combinatorial state automata (CSA) that are more likely to be the minds implemented by brains. Roughly, a CSA is much like a FSA, except that it has a more complex, combinatorial internal structure. Each state is a combination of substates, and any state transition is sensitive to the combinatorial structure of the previous combined state. "CSAs are much less vulnerable to Putnam-style objections than FSAs. Unlike FSA implementations, CSA implementations are required to have complex internal structure and complex dependencies among their parts. For a complex CSA, the right sort of complex structure will be found in very few

physical systems" (1996: 325). Chalmers concludes that brains, but not rocks, implement the complex CSA that is more likely to constitute a mind.

While his points about implementation are well-taken, Chalmers's conclusion that "for a complex CSA, the right sort of complex structure will be found in very few physical systems" does not follow. What does follow is that the proof that the rock implements the functional organization of a thinking organism has to be fixed, but nothing that Chalmers says proves that this cannot be done. As long as no constraints are imposed on the groupings of physical properties that form the implementing states, it is not clear that Putnam's theorem cannot be rehabilitated. As Matthias Scheutz observes, "if no restrictions are imposed on groupings of physical states, then simple, finite, deterministic physical systems…can possibly be seen to implement complex, infinite, and non-deterministic computations" (2001: 551). Indeed, Moore (1990) shows that even a universal Turing machine can be embedded in the motion of a single particle moving in space, bouncing between parabolic and linear mirrors like an ideal billiard ball. The infinite tape and table of instructions can be embedded in the binary development of the coordinates of the particle's position. It might even turn out that if the groupings are defined over the quantum makeup of a rock, the rock implements a complex CSA that is, arguably, constitutive of mind. In light of these results, it is up to the functionalist to demonstrate that there are CSAs implemented by thinking organisms, but by no objects that lack minds.[9]

Furthermore, even if Chalmers is right, and the rock does not implement the functional organization deemed a mind, the other problem – that of my brain's

simultaneously implementing a wide variety of independent CSAs – remains. It is still possible that the same system, e.g., my brain, simultaneously implements many complex independent CSAs, each of which is sufficient to embody an independent mind. Indeed, elsewhere I have demonstrated that even a slight change in the grouping of a single neural property can completely alter the *logical operators* we take the brain to implement. Under a grouping of 0-50mv neural activity into groups of 0-25mv and 25-50mv, the resulting logical operation is AND, but under a grouping into 0-15mv and 15-50mv groups, it is OR. Moreover, it can be shown that even if the proximal sensory-motor I/O organs are given, we can *group* their physical properties in different ways, each matching an implemented abstract CSA (see Shagrir, 2001). Thus the fact that "simple" physical objects such as rocks cannot implement complex CSA does not itself entail that "complex" physical objects such as brains implement a single complex CSA. Indeed, if my brain simultaneously implements different independent CSAs, and if each such implemented CSA is associated with a certain belief-desire scheme (as the functionalist claims), then I must *simultaneously* have all these belief-desire scheme, which I cannot do.[10]

3.3 Why functionalism didn't work

I don't think any of the above arguments deal functionalism a knock-down blow. But they force functionalists to try and escape the conclusions of these arguments by appealing to physical facts, among them, facts pertaining to the distal environment and the

implementing hardware. The appeal to facts in the physical environment is motivated by the need to explain why the thoughts of Oscar and Toscar differ. The appeal to certain conditionals and/or the physics of the implementing hardware is motivated by the need to explain why only humans, but not rocks, implement the automaton that constitutes a mind. Now why is it that this is problematic?

One problem is that the appeal to physical facts blurs the differences between functionalism and classical materialism. The initial attraction of functionalism was the idea that the matter from which we are made is not all that important. What counts is the complexity of the automaton that the organism implements. There could be other organisms, made of very different materials, that implement it as well. It is true that from the beginning, Putnam specified proximal I/O in biological terms. But virtually all the internal causal relations – the whole internal causal network – were specified in some formal-syntactic language. As it now turns out, however, we must also take into account many more physical facts: those pertaining to the distal environment, and those pertaining to the proper implementation of the automaton. We must include many more physical or biological terms in the functionalist specification of mental states.

Worse still, multiple realizability, once the principal argument *for* functionalism, now becomes a threat. Jaegwon Kim, who objected to the multiple realization argument early on (Kim, 1972), pointed out, among other things, that there are seemingly high-order properties, such as the temperature of a gas, that are realized in very different physical substrates, yet are identical with a physical property (mean kinetic energy). Similarly, that a mental property can be realized in both neural tissue and silicon chips

does not entail that this property cannot be identical with one particular physical property. Functionalists have responded that "it is difficult to see how there could be a non-trivial first-order physical property in common to all and only the possible physical realizations of a given Turing-machine state" (Block, 1990: 270-271). It is indeed difficult to see how there could be such a physical property on the weak notion of realization Block has in mind. But, as we just saw, the functionalists cannot use this notion of realization, if they are to avoid the disastrous results of Putnam's theorem. They must use a much stronger notion of realization that will most likely exclude many of the "possible" realizations the weaker notion allows. So there might, after all, be a physical property common to all the realizers of a mental property. This is the physical property to which the functionalists must appeal if they are to avoid the conclusions of the arguments *against* functionalism. It is this result, among others, that inclines many to regard functionalism as no better a hypothesis than classical materialism.[11]

The appeal to physical facts has another drawback, first pointed out by Block with respect to the biological/physical specification of I/O (Block, 1979). The problem is that such specification seems to exclude intelligent beings that lack our biological sensory and motor I/O organs. Such specification is thus unjustly chauvinistic. Why should we assume that there are no possible intelligent beings whose I/O organs differ from ours? If functionalism is correct, then there must be a *non-intentional* equivalence relation over all the biological/physical I/O. But it now turns out that this equivalence relation is to be defined not just over biological/physical proximal I/O, but also over the physical environment, the implementing hardware, and even different abstract automata that

realize the same belief. The functionalist has to take all these issues into account when arguing that two thoughts are of the same type, i.e., have something in common. To accomplish this task, the functionalist must survey of all the beliefs and reasoning – scientific, religious and so forth -- of all humans, individuals and societies, actual and possible. In addition, to satisfy the reductive aspirations of the theory, "this 'something in common' must itself be describable at a physical, or at worst a computational, level" (1988: 100). But it is hard to see how one can succeed in this task. This enterprise, Putnam suggests, is nothing less than Utopian (1994: 510-512).[12]

4. Cognitive science

What can we conclude about the scientific study of the mind from the arguments against functionalism? What is the outlook for the tenets and aspirations of cognitive science, given that functionalism provides its philosophical underpinning? John Searle (1992) contends that cognitive science lacks a firm foundation. Putnam goes further than that, implying that cognitive science is no less than a science fiction (1997, 1999: 118-119). My conclusions are quite different. While I am persuaded that Putnam's arguments against functionalism are by and large correct, I do not think that they pose a threat to cognitive science. Given that cognitive science has generated an impressive empirical and theoretical body of knowledge, I am inclined to reject the other premise in the reasoning: that functionalism provides the conceptual framework for cognitive science. What Putnam's arguments really indicate, in my view, is that computational functionalism is

not helpful for assessing the outlook of cognitive science, and that it rests on a misinterpretation of the scientific and explanatory practices of cognitive science. What follows is a brief diagnosis of the nature of this misinterpretation.

Functionalism became so influential not only because it provides an attractive theory of the mind. Its impact is also due to its linkage with cognitive science. Putnam (1967b: 434-435) even presents functionalism and cognitive science as complementary projects. Functionalism is the project of philosophers whose concern is formulation of a theory of the mental: a comprehensive account of the mind, preferably in non-semantic and non-intentional terms. Inspired by the computational models of cognition, functionalists offer the theory according to which cognitive capacities, thoughts, beliefs and so forth, *are* computational states, specified in terms that are formal-syntactic. Functionalism, however, provides neither a detailed specification of the functional organization of the thinking organism, nor a computational description of the organism's thoughts and beliefs. The functionalist theory, that is, asserts *that* the thought (type) that water is wet is a computational type, but it does not specify *what* computational type it is. This specification is the task of the scientists. It is the aim of the scientific project to specify what type of computational state is each cognitive capacity, thought, and any other mental type. And, according to this picture, cognitive science does exactly that. By specifying the computational structure of cognitive systems, thoughts, beliefs and so forth, cognitive science spells out which computational type is identical with each mental type. Thus functionalism, if correct, provides not only a theory of the mental, but also inspires an overarching picture that explicates the goals and practices of cognitive

science. The aim of cognitive science, on that picture, is to *discover* the functional organization of cognizing organisms, and to specify which computational type is each mental type.

Putnam's arguments against functionalism directly challenge this alleged scientific project, whose aim is to discover the computational types that are identical with mental types. If Putnam's arguments are correct, there are no such identity relations to discover in the first place. The same computational type can yield different thoughts: one is Oscar's thought that water is wet, and the other is Toscar's thought that twater is wet. We could try to enrich our program, and individuate mental content by appealing to communities and environments. But "how useful is it to speak of 'computational-cum-physical states' of such vast systems?" (Putnam 1997: 37). Moreover, even if there were such computational-mental types, it would be impossible to discover them, for the characterization of the content of a thought would require to "describe the content of every belief of every possible kind, or at least every human belief of every possible kind, even of kinds that are not yet invented, or that go with institutions that have not yet come to existence. That is why I say that the idea of such a theory is pure 'science fiction'" (1997: 38).

But there is another option: that the picture inspired by functionalism fails to capture the aims and goals of cognitive science. On this option, cognitive science and functionalism are not complementary projects at all. The aim of cognitive science is not to provide an exhaustive specification of mentality, surely not a full-fledged theory of mental content. Thus Putnam;s arguments against functionalism does not have much

impact on the success of cognitive science. In "Computational psychology and interpretation theory", Putnam (1983b) endorses this option. Arguing that "functionalist psychology" cannot account for mental content, Putnam concludes that "the theory of interpretation and cognitive psychology deal with quite different projects and that to a large extent success in one of these projects is independent of success in the other" (p. 150). Cognitive science seeks to provide a description of how the system of mental representations works; what are the rules of computation that drive the system of representations. Its business is to reveal the interaction between different patterns of representations; to describe, for example, how the visual system extracts information about shape from information about shading, or how it constructs a three-dimensional representation from the disparity between two retinal images. But it does not provide a comprehensive account of the specific content of mental representations, of misrepresentation and so forth. This account comes from elsewhere. This account is the concern of interpretation theory, or what we now call a theory of content. It is interpretation theory, and not the scientific theory, that seeks to provide the pertinent interpretation to the system of mental representations.[13]

I favor the later proposal. I think that it describes much better the objectives and practices of cognitive science. Cognitive science is here to stay for the foreseeable future, and for good reasons. But the question is why Putnam has become so critical of cognitive science in recent years? Why is he presently "convinced that the dream of Psychological Physics that seems to be thinly disguised under many of the programs currently announced for 'cognitive science' will sooner or later be realized to be as illusiory as

Comte's dream of the Social Physics" (1997: 41)? My impression is that the answer has not much to do with the arguments against functionalism, as I think that none of them immediately endangers the programs of cognitive science. To better understand the shift in Putnam's views about cognitive science, we would have to look into Putnam's current Wittgentstenian aversion to the notion of mental representation. But this will have to be dealt with on another occasion.

Notes

[1] For a short survey of Putnam's views, written by Putnam himself, see Putnam (1994b). For a useful survey of functionalism in general, emphasizing the computational version of the thesis, see Block (1995, 1996). For a critical overview of the origins of computational functionalism, and Fodor's contribution to it, see Piccinini (2004).

[2] The page numbers here and throughout the essay refer to the reprinted versions.

[3] Here I assume the truth of the Church-Turing thesis, which states that any input-output function that can be computed by finite means (i.e., by a finite effective procedure) can also be computed by a universal Turing machine. This result motivated Turing (1950), even before Putnam, to associate computing machinery and intelligence.

[4] For a survey of other versions of functionalism, See Block (1979, 1996).

[5] This idea that any token of a mental event is identical with a token of some physical event, but that mental types (properties) are not identical with physical types, is known as non-reductive monism. Putnam advances the view in its most explicit form in 1973 and 1975b. Well-known versions of this view are also advanced by Davidson (1970) and by Fodor (1974). Putnam, however, also notes that "the functional-state hypothesis is *not* incompatible with dualism!" (1967b: 436). Since the hypothesis is simply that mental types are functional types, which are abstract, it is still compatible with the dualistic view that tokens of mental/functional events are not tokens of physical events.

[6] See Block (1996), also for a second-order quantification over the states $S_1,...S_n$.

[7] A probabilistic automaton is a device similar to a Turing machine. The two differ in that: (a) the automaton has a fixed finite memory, whereas the Turing machine has unbounded memory; (b) the state transitions of the automaton might be probabilistic rather than deterministic (though there are also non-determistic Turing machines). All these devices do not have more computational power than a universal Turing machine.

In "The project of artificial intelligence", the first chapter of his *Renewing Philosophy*, Putnam (1992a) qualifies this claim. He explains (pp. 4-7) that he once believed that everything can be seen as some kind of a Turing machine because he assumed that a Turing machine can, in principle, simulate and predict the behavior of any finite system (and a human being is finite in space and time). It was meanwhile proved (Pour-El and Richards 1981), however, that there are possible physical systems, whose time evolution is not describable by a Turing machine computable function, even when the initial condition of the system is so describable.

This result does not defy computationalism, however. For one thing, Pour-El and Richards constructed examples for the wave equation in which the initial data is real recursive (can be sufficiently approximated by a Turing machine), but the solution is not. There is no *empirical* evidence that there are *brain processes* whose evolution is not real recursive (Roger Penrose [1989, 1994] advances a *philosophical* argument that there might be such processes, but Putnam [1995] rightly dismisses it). Secondly, computationalism is not committed to the Turing machine model. It is true that a

universal Turing machine can compute every function that any 'conventional' automaton computes, as well as every function that is 'effectively' computable (assuming the truth of the Church-Turing thesis). But this does not rule out the possibility of devices that compute (non-effectively) functions which are not Turing machine computable. For a detailed relevant discussion, see Copeland (2000).

[8] Putnam himself tends to accept that "*the mind uses a formalized language… both as medium of computation and medium of representations*" (1983b: 141), even while expressing serious reservations about functionalism.

[9] Scheutz (2001) offers an alternative theory of implementation that is relative to a fixed canonical physical theory (e.g., circuit theory), a theory on which the grouping into physical types is already given. In this context, there is a characteristic automaton, which is the most complex implemented automaton. I find the theory interesting and highly useful for the purposes of computer science. But the theory is of little help to the functionalist. As Putnam repeatedly notes, the functionalist has to pick out one physical grouping and not another without appealing to any semantic or intentional traits. Given that functionalism is a reductive theory, it would be unfair to describe humans but not rocks in terms of the pertinent characteristic automaton only because humans are deemed to have minds, and rocks are not. To do so would be totally circular. The computer scientist is in a very different position. To fix the canonical grouping, the computer scientist can and does appeal to traits like goals, purposes, and desiderata such as easy-to-build and user-friendly interfaces. The job of the computer scientist is not, and has never been, akin to that of the functionalist, namely, to provide a reductive theory of content.

[10] Chalmers addresses this possibility, saying that "a given physical hunk of matter can be associated with more than one mind" (1996: 332), yet he does not find it too troubling. But it is troubling. For functionalists have to sort out the 'canonical' from the 'non-canonical' implementations. They have to account in non-intentional terms for the alleged fact that my mind is constituted by one implemented automata as opposed to others.

[11] For recent critical discussions of the multiple realization argument see, e.g., Shagrir (1998), Bechtel and Mundale (1999), Shapiro (2000), and Perebum (2002). It should be noted that there are functionalists who are *also* functional-to-physical reductionists, e.g., Churchland (1984).

[12] Putnam, however, did not abandon all his views about functionalism. He seems to embrace the Aristotelian version even today, and the seeds of his more recent criticisms can be found in his early papers. The chief change in Putnam's views is the total rejection of the reductive assumption taken for granted in the early papers.

[13] There still remains the issue of accounting for the exact relationship between computation and content. This topic has drawn the attention of many philosophers in recent years. For various accounts see Burge (1986), Fodor (1994), Egan (1995), and Shagrir (2001).

References

Bechtel, W. and J. Mundale 1999: "Multiple Realizability Revisited: Linking Cognitive and Neural States", *Philosophy of Science*, 66: 175-207.

Block, N. 1979: "Troubles with Functionalism", in: W. Savage (ed.), *Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science: Volume 9. Minneapolis: University of Minnesota Press, pp. 261-325. Reprinted in Block 1980: 268-305.

Block, N. 1980: *Readings in Philosophy of Psychology*: *Volume 1*, Cambridge Mass: Harvard University Press.

Block, N. 1981: "Psychologism and Behaviorism", *Philosophical Review* 90: 5-43.

Block, N. 1990: "Can the Mind Change the World", in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press.

Block, N. 1995: "The Mind as the Software of the Brain", in D. Osherson, L. Gleitman, S. Kosslyn, E. Smith and S. Sternberg (eds.), *An Invitation to Cognitive Science, Volume 3: Thinking* 2 ed. Cambridge Mass: The MIT Press, pp. 377-425.

Block, N. 1996: "Functionalism", *The Encyclopedia of Philosophy Supplement*, New York: Macmillan.
Block, N. and J.A. Fodor 1972: "What Psychological atates Are Not", *Philosophical Review*, 81: 159-181.

Burge, T. 1986: "Individualism and Psychology", *Philosophical Review*, 95: 3-45.

Carnap, R 1932/33: "Psychology in Physical Language", *Erkenntnis*, 3, 107-142. English version (trans. by George Schick) in A.J. Ayer (ed.), *Logical Positivism*, New York: The Free Press, pp. 165-198.

Chalmers, J. D. 1996: "Does a Rock Implement Every Finite-State Automaton?". *Synthese* 108: 309-333

Chomsky, N. 1957: *Syntactic Structures*, The Hague: Mouton.

Churchland, P. 1984: *Matter and Consciousness*, Cambridge Mass.: The MIT Press.

Churchland, P.S. and Sejnowski, T. 1992: *The Computational Brain*. Cambridge, MA.: MIT Press.

Copeland, B.J. 2000: "Narrow Versus Wide Mechanism", *Journal of Philosophy*, 97: 1-32.

Davidson, D. 1970: "Mental Events", in L. Foster and J.W. Swanson (eds.), *Experience and Theory*. Amherst: University of Massachusetts Press, pp. 79-l0l.

Descartes, R. 1637: *Discourse on the Method*. In J. Cottingham, R. Stoothoff and D. Murdoch (trans.), *The Philosophical Writings of Descartes: Volume 1* (1985). Cambrige: Cambridge University Press.

Egan, F. 1995: "Computation and Content". *Philosophical Review* 104: 181-204.

Feigl, H. 1958: "The 'Mental' and the 'Physical'", in H. Feigl, M. Scriven and G. Maxwell (eds.), *Concepts, Theories and the Mind-Body Problem*. Minnesota Studies in the Philosophy of Science: Vol. 2. Minneapolis: University of Minnesota Press, pp. 370-497. Reissued in 1967 with a postscript by University of Minnesota Press.

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

Fodor, J.A. 1974: "Special Sciences, or the Disunity of Science as a Working Hypothesis", *Synthese*, 28: 97-115.

Fodor, J.A. 1975: *The Language of Thought*, New York: Thomas Y. Crowell.

Fodor, J.A. 1994: *The Elm and the Expert, Mentalese and its Semantics*, Cambridge, MA.: MIT Press.

Fodor, J.A. and E. Lepore 1992: *Holism: A Shopper's Guide*, Oxford: Blackwell.

Harman, G. 1988: "Wide Functionalism", in S. Schiffer and S. Steele (eds.), *Cognition and Representation. Boulder*: Westview, pp. 11–20.

Hempel, C.G. 1949: "The Logical Analysis of Psychology", in H. Feigl and W. Sellars (eds.), *Readings in Philosophical Analysis*, New York: Appleton-Century-Crofts., pp. 373-384.

Kim, J. 1972: "Phenomenal Properties, Psychophysical Laws, and the Identity Theory", *The Monist*, 56: 177-192.Moore, C. 1990: "Unpredictability and Undecidability in Dynamical Systems", *Physical Review Letters*, 64: 2354-2357.

Penrose, R. 1989: *The Emperor's New Mind*. Oxford: Oxford University Press.

Penrose, R. 1994: *Shadows of the Mind*. New York and Oxford: Oxford University Press.

Pereboom, D. 2002: "Robust Nonreductive Materialism", *Journal of Philosophy*, 99: 499-531.
Piccinini, G. 2004: "Functionalism, computationalism, and Mental States", *Studies in the History and Philosophy of Science*, 35: 811-833.

Place, U.T. 1956: "Is Consciousness a Brain Process?", *British Journal of Psychology*, 47: 44-50.
Pour-El, M.B. and Richards, I. 1981: "The Wave Equation with Computable Initial Data such that its Unique Solution is not Computable", *Advances in Mathematics*, 39: 215-239.
Putnam, H. 1960: "Minds and Machines", in S. Hook (ed.), *Dimensions of Mind*, New York: University of New York Press, pp. 148-80. Reprinted in Putnam, H. 1975a: 362-385.
Putnam, H. 1963: "Brains and Behavior", in R. Butler (ed.), *Analytical Philosophy. Second Series*, Oxford: Basil Blackwell & Mott, pp. 1-19. Reprinted in Putnam, H. 1975a: 325-341.
Putnam, H. 1964: "Robots: Machines or Artificially Created Life?", *Journal of Philosophy*, vol. 61: 668-691. Reprinted in Putnam, H. 1975a: 386-407.
Putnam, H. 1967a: "The Mental Life of some Machines", in Hector-Neri Castañeda (ed.), *Intentionality, Minds and Perception*, Detroit: Wayne State University Press, pp. 177-200. Reprinted in Putnam, H. 1975a: 408-428.
Putnam, H. 1967b: "The Nature of Mental States" (originally published as "Psychological Predicates"), in Captain, W. H. and Merrill, D. D. (eds.), *Art, Mind and Religion*, Pittsburgh: University of Pittsburgh Press, pp. 37-48. Reprinted in Putnam, H. 1975a: pp. 429-440.
Putnam, H. 1973: "Reductionism and the Nature of Psychology", *Cognition*, 2: 131-149. Reprinted in Putnam, H. 1994a: 428-440.
Putnam, H. 1975a: *Mind, Language and Reality, Philosophical Papers,Volume 2*, Cambridge: Cambridge University Press.
Putnam, H. 1975b: "Philosophy and Our Mental Life", in Putnam 1975a: 291-303.
Putnam, H. 1975c: "The Meaning of 'Meaning'", in K. Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII, Minneapolis: University of Minnesota Press, pp. 131-193. Reprinted in Putnam 1975a: 215-271.
Putnam, H. 1980: "Models and Reality", *Journal of Symbolic Logic*, 45: 464-482. Reprinted in Putnam 1983a: 1-25.
Putnam, H. 1983a: *Realism and Reason, Philosophical Papers, Volume 3*, Cambridge: Cambridge University Press.
Putnam, H. 1983b: "Computational Psychology and Interpretation Theory", in Putnam 1983a: 139-154.
Putnam, H. 1988: *Representation and Reality*, Cambridge, Mass.: The MIT Press.
Putnam, H. 1992a: *Renewing Philosophy*. Cambridge: Harvard University Press.
Putnam, H. 1992b: "Why Functionalism didn't Work", in J. Earman (ed.), *Inference, Explanation and Other Philosophical Frustrations*, Berkeley: University of California Press, pp. 255-270. Reprinted in Putnam, H. 1994a: 441-459.
Putnam, H. 1994a: *Words and Life*, edited by J. Conant, Cambridge, Mass.: Harvard University Press.
Putnam, H. 1994b: "Putnam, Hilary", in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Cambridge: Blackwell, pp. 507-513.
Putnam, H. 1995: "Review of Roger Penrose, *Shadows of the Mind*", *Bulletin of the American Mathematical Society* 32.3: 370-373.
Putnam, H. 1997: "Functionalism: Cognitive Science or Science Fiction? ", in Johnson, D.M. and Erneling, C.E. (eds.), *The Future of the Cognitive Revolution*, Oxford: Oxford University Press, pp. 32-44.
Putnam, H. 1999: *The Threefold Cord:Mind, Body, and World*, New York: Columbia University Press.
Ryle, G. 1949: *The Concept of Mind*, London: Hutchinsom.
Scheutz, M. 2001: "Causal vs. Computational Complexity?", *Minds and Machines*, 11: 534-566.
Searle, J. 1992. *The Rediscovery of the Mind*, Cambridge: MIT
Shagrir, O. 1988: "Multiple Realization, Computation and the Taxonomy of Psychological States", *Synthese*, 114: 445-461.
Shagrir, O. 2001: "Content, Computation and Externalism", *Mind*, 110: 369-400.

Shapiro, L. 2000: “Multiple Realizations”, *Journal of Philosophy*, 97: 635-654.
Smart, J.J.C. 1959: “Sensations and brain processes”, *Philosophical Review*, 68: 141-156.
Turing, A.M. 1936: “On Computable Numbers, with an Application to the Entscheidungsproblem”, *Proceedings of the London Mathematical Society* (2), 42: 230-265. A correction in 43 (1937): 544-546.
Turing, A.M. 1950: “Computing Machines and Intelligence”, *Mind,* 59: 433-460.