How Would You Know if You Synthesized a Thinking Thing?

MICHAEL KARY[1] and MARTIN MAHNER[2]
[1]*Chaire en Gestion des Bioindustries, Université du Québec à Montréal, and Department of Mathematics, Boston University, Boston, MA, U.S.A.*
[2]*Center for Inquiry — Europe, Rossdorf, Germany; E-mail: mahner@gwup.org*

Abstract. We confront the following popular views: that mind or life are algorithms; that thinking, or more generally any process other than computation, is computation; that anything other than a working brain can have thoughts; that anything other than a biological organism can be alive; that form and function are independent of matter; that sufficiently accurate simulations are just as genuine as the real things they imitate; and that the Turing test is either a necessary or sufficient or scientific procedure for evaluating whether or not an entity is intelligent. Drawing on the distinction between activities and tasks, and the fundamental scientific principles of ontological lawfulness, epistemological realism, and methodological skepticism, we argue for traditional scientific materialism of the emergentist kind in opposition to the functionalism, behaviourism, tacit idealism, and merely decorative materialism of the artificial intelligence and artificial life communities.

Key words: artificial intelligence, artificial life, biomimetic computation, cognitivism, emergence, functionalism, lawfulness, life, materialism, mind, Turing test

1. Introduction

In daily life we conveniently ignore the premise underlying this journal, namely that thoughts need not be exclusively the function of brains. We refer to this as a convenience, but surely that understates the matter: imagine for a moment what life would be like otherwise. Imagine for a moment that we could no longer take for granted that a thing is thoughtless, for no other reason than that it is brainless. Imagine if we could not be assured that our creations were not sentient, simply because we could not endow them with a nervous system. Imagine the difficulties we might face if we took seriously the implication of artificial intelligence, to judge every artifact on its own merits, without the ready answers supplied by neurocentrism.

For one, how would we evaluate our garage tinkerings? Might we one day fit together a right combination of gadgets and contraptions, one that gave a thing with high-level thought processes, and hence perhaps in some sense, a person? If so, we might not notice, and through further tinkering, unwittingly slay it. Or might we, at our desks writing computer programs, happen to implement some basic algorithm of thinking, if such a thing exists? We might not become aware of our achievement, and then inadvertently erase it, and so again have committed an atrocity. And would not the open-minded among us, at work in their kitchens on a new recipe for breakfast, have to contend with the possibility that they might just stumble upon

Minds and Machines **12:** 61–86, 2002.

the magic combination of ingredients, cooked in just the right way, that would result in a thinking omelette — one with hopes, dreams, and aspirations? If man in the information-age kitchen is to stand against neuro-chauvinism in matters of the mental, he must find some other way of recognizing psychological endowments, so that he will neither carelessly eat any of his possible omelette-person creations, nor starve to death himself. After all, even if we might allow that man might one day be genius enough to create a sentient omelette, we should not expect him to be so lucky as to cook one up that could also speak English.

Even though most of us take for granted that we will never be brought up on capital charges for tinkering in our garages, aborting computer programs or eating an omelette, such possible ends must be very real to functionalists and cognitivists, that is, to those for whom matter does not matter, but only the pattern of what it does or how it is organized (Putnam, 1975; Dennett, 1978; Block, 1980a; Pylyshyn, 1984; Langton, 1989, 1991; Boden, 1996). Citings of putatively absurd examples of things which, if functionalism were true, ought to be investigated for mental abilities, are seen by some anti-functionalists as refutations in themselves. We do not share this view, since committed functionalists already allow for such extremes. As Putnam (1975, p. 291) affirmed[1]: "We could be made out of Swiss cheese and it wouldn't matter".

Such is the philosophy to which proponents of the strong artificial intelligence (AI) and artificial life (AL) programmes adhere, and it has extraordinary consequences for daily life. For if mental processes are properly characterized only in terms of some pattern of computation, or in some other abstractly formal or functional way, then we must respect the possibilities for sentience within all things whose internal activities are not understood from this perspective. Since functionalists are yet to provide any further specifications for this characterization, every thing with any internal activity of any kind falls into this category, including, for example, omelettes. Who can imagine the complicated reactions and changes of molecular state going on when cooking a complex omelette, which might somehow be equivalent formally, to some exotic thinking algorithm? This is not a joke: if in serious articles in serious journals the question can be raised "whether in well-known reaction mechanisms computational functions occur, and, if they do, what role and purpose they serve" (Arkin and Ross, 1994); and if we are to understand that, as these authors tell us, there are "computational functions in glycolysis", i.e. that in addition to releasing energy, a glycolytic reaction is also a tiny computer that nature uses to figure out something or other, then what is going on at levels deeper still? Who knows even within one single atom, what complicated forms might be 'computed' by the dynamics of the quarks making up the atoms, or whatever there is, if anything in turn, making up quarks. How can we know even whether it is not the case that every single atom thinks, feels, and dreams?

These questions have proved troublesome for many. While on the one hand computer enthusiasts have been insisting for half a century or more that artificial intelligence and even artificial life are either already here[2] or nearly so, on the other

it is still the case that many otherwise educated people, and even some scientists, do not accept that birds or dogs or even apes might have minds. And it has not been long at all since the time when esteemed individuals could publicly discount the cognitive and affective capacities of the members of one ethnic group or another, or of one sex or the other. Finally, ever since at least the time of Newton and Laplace, a few thinkers — they ought to object to that appellation — have denied the reality of independent thought altogether, claiming that the activities of the brain are no more intelligent, creative, or wilful than those of a pendulum clock.

2. Identifying Activities and Roles

How then do we know whether or not a given thing thinks? How do we know whether or not we do, when awake or when asleep? For that matter, how do we know whether or not a so-called computer computes, whether a bird flies (Buschlinger et al., 1998), or anything of the sort?

Some of these are functions in the sense of activities in and of themselves, and some are functions in the sense of either roles or tasks in a larger process. While 'function' is traditionally a problematic term in philosophy, used to name at least six concepts (Walsh and Ariew, 1996; Mahner and Bunge, 2001), these are the only two we need here, and they can be defined technically and precisely (Mahner and Bunge, 1997, p. 155ff, 2000, 2001; or from a slightly different and more computational perspective, Chandrasekaran and Josephson, 1997). Regardless, for our purposes it is enough to clarify them as follows: the activities of a thing are its changes of state, more specifically the totality of all its processes, while its role within a larger process consists of all its effects on any of the other things involved in that process. Both concepts are non-teleological; if useful we can further specify that its tasks are those roles for which it has been designed, but often the two are used synonymously, or at least without the intention of emphasizing the distinction. Example 1: two of the activities of an internal combustion engine are the combustion of fuel and the rotation of the crankshaft. Two of its tasks in the operation of a car are the driving of the wheels and the alternator; while an additional role is to contribute to acid rain. Example 2: one of the activities of the sun is thermonuclear fusion. One of its roles is the warming of the planets. Example 3: one of the activities of the heart is rhythmic contracting; one of its roles is to pump blood; another is to make sounds. Example 4, building on an example given by Chandrasekaran and Josephson (1997): the activities of a well-constructed wooden chair are close to nil: it just sits there. Two of its tasks are to allow a person to just sit there on top of it, and to be decorative.

However precisely defined or in classification further refined, the distinction between activities and tasks is crucial, for tasks and roles may, if we are lucky, be accomplished by any of several different things operating in different ways, while activities simply are what they are. Accordingly, which are which may be controversial. For example, in our view, thinking, being simply a brain process, is

to begin with an activity in and of itself, while computation is in the end a task. That is to say the activities of a computer, and possibly even of our own neural subsystems responsible for calculation, are really computations only in terms of a larger process involving higher cortical functions, wherein the tokens, symbols and actions are interpreted as ideas, in particular mathematical ones. Yet dualists of a certain kind might consider every activity of the brain to be thoughtless if not part of a larger process involving an immaterial mind, while cognitivists of a certain kind might consider computation an intrinsic activity. As Sloman (1996) has convincingly shown though, the latter view runs into trouble when trying to actually define computation in stand-alone terms, i.e., without recourse to the concepts of symbol, meaning and the like (symbolize what for whom?); whereas a definition of computation in terms of Turing equivalence is inadequate for other reasons as a foundation for cognitivism. In our view even the concept of computation as what can be generated by a Turing machine is not adequately independent of a larger context, since, for one thing, Turing's theory is about 'purely formal structures' (Sloman, 1996), which brings up the problem of what precisely a formal structure is if not an idea. Nevertheless, as interesting as the question of whether or not a computer computes may be, it is not our subject: we are concerned here exclusively with identifying mental activities[3].

The outline of the procedure is not complicated (for the contrary view see Buschlinger et al., 1998). In order to decide whether or not any thing is doing anything, we require some sort of criterion. Example: a substance acts as an acid if it turns litmus paper red. The best criteria are grounded in definitions of concepts representing the activity in question, and the best definitions are those which have integral places within well-confirmed scientific theories, for having a key role within a coherent and largely true body of knowledge. For example, the definition of an acid as a proton donor, integrated into a larger understanding of chemistry, including the spectrochemistry of organic compounds, can explain why acids turn litmus paper red. This adds necessary depth and significance to both the criterion and the definition. In turn the latter, with the aid of sophisticated chemistry, can be used as a criterion more refined than the litmus test.

But where do we begin? We begin as always in the only possible place, namely wherever we are. In other words, we begin, as in every problem, with our background knowledge and preconceptions. We must begin but need not end there: we can, if we find it worthwhile, proceed to investigate. We do so by hypothesizing new definitions of old concepts from our background knowledge, new concepts, new trans-phenomenal entities and processes, new relationships between them, and testing for the consequences direct and indirect. Eventually, we may come back to correct or even discard many of our original presuppositions, and sometimes even the original problem. In short, we use the scientific method.

Let us start with a specific example of a mental function, namely memory. Somewhere along the way in this process of inquiry we might be able to abstract, to a greater or lesser extent, away from our understanding of specific things that

remember and propose a general definition of memory, one which hypothesizes a characteristic or characteristics common to all known instances of memory. Depending on the context, as such it either predicts or defines that to hold for all such instances to be found in the future. Depending on its results and coherence, it may lead us to reject a small number of things accepted as examples of memory in the past. Thus even though a definition within the context of the corresponding theory, it begins and ends life as a hypothesis. Eventually one may hypothesize something either slightly different or more or less inclusive, as was done several times in the theory of acidity: for example, an acid may alternatively be defined as an electron-pair acceptor.

In the case of memory, one does not have to go along very far before being able to put forth a preliminary definition such as the following:

> Thing *A* has memory of fact *B* if, directly or indirectly, *B* has left some identifiable trace on *A*.

This is not offered as quintessential or even complete ('identifiable trace' has to be defined in terms of a lawful relationship between the state of A, and B; see Kary, 1990), but merely as an outline of a proposal to be examined for the time being.

We can then use this as a criterion to decide whether things other than the brain, or even the brain itself, can remember. We find that according to this criterion, many things can and do have memory; and actually that the human memory system is rather poor at remembering, because, among other problems, our memories are often altered by things unrelated, the more so as time goes by.

Note though that unlike the definition of acidity, this definition specifies no common mechanism. Thus while it allows for things such as various computer storage devices, blackboards, metals, and so on to all have memory, it means that studying any one of them will guarantee us no knowledge whatsoever of any of the others, and none again of mental memory. Likewise, without a mechanism, the definition itself is not specific enough to tell us anything about any of them in particular. At best, it just helps guide our study of mental memory, or helps tell us what to look for when trying to design and evaluate remembering devices. Put another way, it defines memory in terms of a functional role, and not in terms of the functional activity of mental memory. For example, we know our human memory systems do many things in addition to satisfying the above, and that many of them are integral to our experience of memory, yet may interfere with the realization of the memory criterion. This is so much the case that upon further study, we might decide that this definition is inadequate to characterize even the role of mental memory, and so choose to rename the concept defined above as something like 'information' (Kary, 1990). This might nevertheless still be a component of some eventual redefinition of memory, as reserved for something closer to the functions more specific to the mental memory system. Or, we might just as well keep this as the fundamental core of a definition of memory, and recognize that it is but one of the functions of what we traditionally think of as our own memory systems. The two avenues lead to competing or complementary theories, to be decided upon, or alternated between,

in the usual way: on the basis of their overall results, explanatory power, ease of use, coherence, and fit with the rest of scientific knowledge.

3. Original and Simulation

Before moving on to consider any more advanced cognitive or emotional abilities, let us apply the same procedure to an entirely different and much simpler biological function often brought up in such discussions, namely digestion. For example, consider the following possible definition:

> For thing *A* to digest things of kind *K* is for *A* to break down chemically things of kind *K*, and store the liberated energy in bonds of molecules of some other type.

Again, this is a starting point for an argument, not an ending point.

A first question: are there artificial digesting things? The answer seems to be yes: according to the above definition, many a chemical factory qualifies as such. And in any case there are semi-artificial digesting things, such as breweries. Now consider the following question. Suppose we were to develop an excellent mathematical model of some digestive process, or even a more general theory of all digestive processes, which we could use to model any individual one. Suppose further that we implement such a model on a computer, either by solving it completely or by using it to run a simulation of a digestive process, perhaps even complete with eye-catching animations graphically depicting the entire affair. Is our computer then digesting?

Of course not. We know that it is not digesting by virtue of the fact that such a simulation does not satisfy the digestion criterion: no chemical reactions of the sort required occur. Electronic computers are for all intents and purposes chemically inactive. Their inner machinations, no matter how advanced the programming, are changes of magnetic polarity, electron flows, phosphorescence and the like, not chemical decomposition and recomposition. We are in accord then with Harnad's analysis (1991), and comparable analyses by Sober (1991), Pattee (1989), and others: a computer model of a fire is not hot.

The reaction of some cognitivists, upon hearing this devastating news, underlies the recursive nature of their philosophy. They concede that while it is true that a computer model of a fire is not hot, it should contain a *computer model* of heat, as if that somehow solved the problem. The idea that it does comes in its specifics from a historical digression, which we outline below; and in general from something more pervasive, which we return to address in Section 6.

Dretske (1985) claimed that since computer simulations of hurricanes do not blow down trees, computer simulations of problems do not solve them. Rapaport (1988) responded that since such simulated solutions of simulated problems really do solve the originals, then ... well, let us say then that Rapaport concluded that whether computer simulations of hurricanes blow down trees or not is irrelevant.

But while we agree with Dretske's premise, we find his conclusion unrelated. Hurricanes and trees are concrete objects with substantive properties satisfying natural laws. Problems and solutions, and more generally, the concepts and propositions that they consist of, are abstract objects with conceptual properties attributed by hypothesis, and which we agree to manipulate according to rules. We can designate these abstract objects by symbols or tokens, and manipulate the symbols or tokens in thought, via pencil and paper, or by machine or even human proxy. When we know what we are doing, they represent for us, but not to the pencil or paper or machine, and sometimes not to the human deputy, the corresponding concepts; and, appropriately implemented, the ensuing results represent to us solutions of corresponding problems. Simulation of a concrete process requires though in addition to all these steps, the formation of conceptual or symbolic models of physical things, their properties and laws. What the computer can or cannot do with regard to solving problems, and Rapaport's arguments, are thus irrelevant to what it can or cannot do with regard to instantiating substantive properties, such as being alive or thinking.

All this is a distraction. None of it solves the real problem: the fact is that for the committed functionalist, conceding that a computer model of a fire is not hot amounts to philosophical suicide (see below), while refusing it leads to actual suicide. Indeed, if it were otherwise, why not then reduce the ecological burden on the planet, and subsist on computer models of food cooked on computer models of fire? This plan must have already occurred to some in the AL community, who claim for example that what we eat are patterns, not things (Grand, 2000, p. 56). Idealists, old and new alike, always seem to run headlong into the problem of philosophizing on an empty stomach.

Thus it is that the computer gives us only a simulation, which we may find only more or less useful, in our study of the original. Neural network simulations, for example, have been of use to neuroscience. (Perhaps they will be, or have already been, of greater use to computer science.)

Let us turn then to some key mental process, such as thinking, or even to the process of life itself. Suppose that we were able to develop excellent mathematical models of all thinking processes, or even of life itself, and suppose we could implement those models on a computer. Suppose that we could similarly realize them for individual persons, such as the present authors, in such a way that the computer could model our entire past and future behaviour, and even replace us at any cocktail party, or for that matter, write the rest of this article. Surely these simulations could contribute to our scientific, or at least medical knowledge — even though we would have needed extensive neuroscientific and biological knowledge to construct and validate them in the first place — but could the computer running them be said to be thinking, or alive? Need we worry about violating its personhood by destroying it, or even simply turning it off?

Of course not, and for the same reason. Just as it must eventually be conceded — one way or the other — that a computer model of a fire is not hot, it must

then be conceded that a computer model of a living thing is not alive, a computer model of a thinking thing is thoughtless, and a computer model of a person is not a person. Modelling what a thing will do does not constitute being or existing as that thing. If it were to, then Kepler would be a solar system and Galilei a cannonball. Such is what follows from the central claim of the strong artificial life and artificial intelligence programmes, that correctly implemented simulations are as genuine as the things they imitate (Langton, 1989, p. 33). Indeed, since it is possible to 'simulate' things which do not exist, such as a geocentric solar system, believers in AI and AL must grapple with the ghostly problem of deciding what exactly a genuine non-existing thing could be (Mahner and Bunge, 1997, p. 152).

Thus to be alive or to think, a thing has to satisfy, and not merely simulate, whatever it is that characterizes thinking or living things. Still, the computer faithful might object that, since we do not yet know what an appropriate general characterization of a thinking thing is, we cannot yet say that an appropriately running computer might not itself one day satisfy this characterization, and not merely simulate it. And yes, perhaps nearly one century ago, it might have been possible to uphold this view. At that time, it was thought that the sole distinguishing characteristic of brain processes was the firing of electrical signals by the neurons in an all or none fashion. So why not construct a brain out of man-made electrical relays, which would serve as synthetic neurons? All one would have to do then would be to replicate the firing pattern found in some individual thinker, and one would have synthesized a thinking thing.

Such claims would be true if that were all there were to neurophysiology. However, we have known for a long time now that brains are far more than systems of electrical relays, although they are that too. Neurons show spontaneous activity while relays do not, and thoughts involve cascades of chemical and biochemical and biological processes, such as neurotransmitter uptake and release and dendritic growth. These are known to be part of thought because specific psychological characteristics and disorders have been identified with these mechanisms — learning with the formation of new neuronal supersystems, by the mechanisms of dendritic sprouting and the growth of new nerve cells; schizophrenia with the details of neurotransmitter release and uptake across the synapse; and presumably, genuine spontaneity and creativity with various genuinely spontaneous processes of neuronal activity.

Faced with such facts, the proponents of artificial intelligence and artificial life may only propose the construction of some more realistic silicon or other element that would make for a more believable synthetic neuron. After all, their basic claim is that matter does not matter, as long as the same functions are displayed. Thus the next question: can in fact one outline classes of properties and processes in a general, matter-free manner? If not, why not, and if so, when?

4. Substance and Function I

Again, let us begin with a much simpler example. Consider the process of rolling along a surface. What things can roll? Anything that is round or close to it. This requirement is quite general but not completely stuff-free: for example, since the shape is specified, we know that only solid objects can roll along a surface for any extended length. Some fairly cohesive liquid or even gaseous systems might roll or come close to it though, at least for short distances.

Now consider a slightly more demanding process, namely what one might call 'wheeling'. Recall that the difference between what one might in general call a 'roller', namely anything with a round cross-section, and a wheel, is the presence of a hub and axle. We find that, while solids, liquids and even gases might roll, and that while rollers do occur frequently in nature, only solids can be used to build wheels, and no wheels occur naturally. Furthermore, not every combination of solids makes for a good wheel: for one to be used in a simple pushcart, for example, the materials must be chosen so that among other things, the friction between the outer rim and the ground is always appropriately greater than the friction between the hub and the axle. Thus, not only are there new restrictions placed on the possible components, but also upon the internal structure of the system (the friction at the hub), the external structure (the friction between the rim and the ground), and even upon the relationship between the internal and external structures.

The point of this example is to show that the more complex and specialized the function, the more it becomes tied to the special properties of specific materials and systems. This is only a general rule of thumb; sometimes even very simple properties are tied very specifically to special systems. Consider for example the atomic property of being able to join together with like atoms to form long chains and branched molecular systems: there is but one atom that has this property, namely the carbon atom. There is a simple molecule that has a similar property, namely SiO, silicone; but the dissociation energy of such bonds between SiO molecules is significantly greater than that between carbon atoms, so only carbon is suitable as a biomolecular building block for the temperatures encountered here on planet Earth.

It is then simply a reflection of the facts of nature that while some process classes are very large, others are extremely small. In the case of the class of thinking processes, its only known members from which to base a definition on are certain special processes that occur in the brains of higher vertebrates, and possibly animals of some other kinds, such as octopodes. In the case of the more specialized thought processes of consciousness, self-awareness and so on, there is a still narrower category whose exact membership remains the subject of empirical investigation. In the case of life processes, there is a broader category that includes all systems which are simultaneously capable of undergoing processes such as genetically templated protein synthesis, certain trans-membrane active transport processes, certain energy conversion processes, and so on (Mahner and Bunge,

1997, 2000; Boden, 1999). Any definition of life or thought that goes beyond these classes to allow materials and systems which are not capable of performing all these activities, without proving at least one characteristic in common with them and explaining why and proving that it, and only it, is essential, is by definition *ad hoc*, and thus amounts to no more than creating life and mind by decree (cf. Mahner and Bunge, 1997, and also Section 5).

Yet wait. To return to one of our original examples, what if one day an enterprising engineer with too much time on his hands were to develop, against all odds, an omelette–human interface, complete with universal translator, that would allow us to hear, to our surprise and consternation, an omelette's heretofore silent screams. Faced with this 'reality', would anti-functionalists not then look foolish? Would we be able to turn a deaf ear to some omelette's eloquent and dignified defense of its existential rights, and callously eat it nonetheless, proclaiming that science precludes the possibility of its sentience, no matter how convincing the drama? Or, what most would consider a more realistic proposition, a comparable situation for a human-looking robot with a computer chip for a brain. Is not then the ultimate criterion of mentality functional, or else behavioural, and not material and mechanismic? Or more accurately, should not our ultimate definitions of mentality be in functional or else behavioural terms, and not material and mechanismic ones?

To begin with, it depends on what is meant by a functional definition. The idea needs to be split at least three ways, which we examine in this and the following two sections. The first is that of a definition given only in terms of abstract patterns and mathematical relationships, without reference to material things and material properties; the next, that of a definition in terms of the functional role of a thing, as opposed to its activities; and the third, as described by, e.g., Block (1980a), that mental states should be defined in terms of behavioural items plus other mental states, and the causal relations between them.

As for behavioural, i.e., superficial, definitions (see also Section 7), they may serve as starting points but not as ending points. If it were otherwise, they might save the pre-wired sentient omelette, but not the sentient victim of amyelotrophic lateral sclerosis not so equipped; nor explain why we should go to the effort to outfit either one, or why one but not the other. Nor could a definition of thought as some sort of pattern, devoid of substance, help: because there is always, at smaller or larger scales, over different combinations of components, over longer and longer time frames, involving any or all possible properties, one more possible pattern, one can never know whether somewhere or other in some as yet undiscovered fashion, the magic one is not hidden. Nor can a functional definition explain, without reference to things and their properties, why one pattern should be a thought and another not. Only the mechanismic, matter-based definition is explanatory: ALS attacks the motor neurons but not the cerebral cortex, while omelettes, no less than human-looking robots, do not have any neurons at all.

The fact is that we do not live in a world where properties either exist without matter, or are unconstrained by the particular nature of the matter in question — be-

cause the nature of a thing is precisely its properties[4] and their lawful relationships. We no longer, if anyone ever did, live in a world where completely ungrounded philosophical speculation — in this context, biology without reference to organisms, and psychology without reference to thinking organisms — can be taken seriously. And unlike a century ago, we nowadays know enough about the brain to know that thinking semiconductor and metal chips (as opposed to ones which merely simulate one or more aspects of thinking) fall squarely into this category no less than thinking omelettes. No doubt a robot might fake humanity or animality well enough to stop us from destroying it; hardly an achievement though, since even mutilating a doll is disconcerting enough for most non-psychopaths. This shows how superficial appearances may be.

And regardless of whether or not it could fantastically turn out that with enough luck, ingenuity and funding, we could make a brain out of eggs, or silicon and copper, or a silk purse out of a sow's ear, this achievement would not make functionalism any truer. Life and thought would not become properties of the organization of things, but remain properties of the things that are so organized. Definitions of patterns without things are not well-formed. For example, it will not do to try to define memory elliptically, by saying that a memory is an identifiable trace, hoping not to mention the things that are so traced, because in fact the relationship involved is predicated of things and events, and events themselves are changes in things. There is no such thing as a change in nothing. Even in mathematics, a relationship cannot be defined without the use of relata. Nor is it possible to define memory, life, thought, or any other material property using only mathematical relata, i.e. as "mind, life or whatever is (some relationship between or collection of mathematical objects)", if only because, to choose two amongst a nearly limitless array of reasons, that would require (a) living things to be composed of mathematical objects; and (b) mind to have originated prior to or simultaneously with life, since mathematical objects are conceptual. Both of these results are counterfactual, the last obviously and the first, if to some not obviously, because otherwise all the conservation laws would be false, since mathematical objects are hypothesized into existence. Surely the next option, that living things are not real, is no better. Finally, it is not acceptable either to say that "a thinking, living, or whatever thing is a material system having (some mathematical property)", because material systems do not have any mathematical properties, they have only material ones. These may be expressed with the aid of mathematics, so that they look a little like mathematical properties; but they are not.

Thus while Boden (1999, p. 231) says that apart from metabolism, various other properties of life, such as the ability to grow or propagate, can arguably be defined functionally, we argue that they cannot. Consider the simplest example, growth. It may appear that a growing thing could be defined purely in terms of an abstract pattern, by saying "growth is an increase in quantity" — until one tries to define the concept of an increase in quantity. It has to be defined at minimum something like this:

> Let F be a reference frame (incidentally, unlike a coordinate system, a material object), and U a system of units including a mass unit and a time unit (both of which in turn must be referred to specific material systems, such as a specific platinum–iridium bar, or a caesium atom transiting between two states). Then for an object O to increase in quantity over a time interval Δt is for its mass (in the appropriate units) at time t_2 with respect to frame F to be greater than its corresponding mass at time t_1, where $t_2 - t_1 = \Delta t$.

So, while some definitions referring to material systems appear to be in terms of abstract patterns, that is only because they have not been analyzed down to their foundations. The question is not whether thought and life can be defined in a stuff-free manner, but simply what the material scope of the material properties will be. We return to this question in Section 5, but for the moment, regarding the prospects for making synthetic neurons out of any other materials, including new ones not yet invented, functionalists do not appreciate that we are often barely able or unable to get the functions we need in the most mundane of technological endeavours, let alone to recreate something so staggeringly complex as nervous tissue. Just to find a material to make filaments for light bulbs was a historic struggle, and we are yet to find a material with the right combination of properties to satisfactorily pave roads in harsh climates. Or consider the problem of making a material suitable for high quality optical lenses: if the molecules comprising it are too large, as is the case with all plastics, points of light will be transmitted with halos surrounding them. Thus, the molecules must be small. But since there are only a finite number of elements, there are only a finite number of sufficiently small molecules that could possibly be candidates for lens material for this reason alone. And as luck would have it, there are only two or three materials that do fit the bill, and in the end, only one of these, glass, that has all the properties needed to make practical lenses. Likewise, there is every reason to affirm that there must be molecular-scale mechanisms at the root of life and thought; not just because chemical reactions by definition operate only at that scale, but also because of the vast changes in physical constraints that occur as the size scale is increased, such as the emergence of friction. Since there are only a finite and small number of elements, there are only a large but still very finite number of possible molecules out of which to even conceivably build synthetic neurons, neurotransmitters, metabolic systems and the like. The vast majority of these will be unsuitable for elementary reasons alone, such as chemical instability or impossibility. Nearly all of the remaining will be unsuitable for other elementary reasons, such as being quasi-inert. When all is said and done we will be lucky if there is a single molecule, let alone organelle or cell, fundamentally different from what is already in place, yet which could be used to substitute for any but a few minor subtasks.

By contrast, an ability to be controllably switched between two easily recognizable states can be easily engineered. That is why the modern computer is designed the way it is, and why it allows so much flexibility in choice of materials. For example, computer memory has over the years been manufactured from paper,

tiny magnets on wires, and rust, among other things. It may even be possible to build practical computers based on chemical reactions (Arkin and Ross, 1994) or quantum delocalization (known today by the egregiously deliberate misnomer quantum 'teleportation'). No doubt this unusual, though still not unlimited, adaptability of the processes within their own domain of expertise has encouraged some computer scientists to over-generalize the same adaptability to the far more complex processes of thought and life.

In other words, whatever our success in understanding the brain, its level of complexity is such that it is unreasonable to assume that we will ever be able to substantially duplicate — as opposed to simulate — more than a very small number of its myriad activities in anything substantially different from it (see also Section 8). The development of a minor prosthesis for some specialized function though, as is well along for the transfer of auditory and visual signals, is at least a reasonable quest. So are attempts to handle at least some tasks that we accomplish intelligently via devices or algorithms which are not intelligent, as have often succeeded in the past. Likewise, attempts by devices or algorithms which, though just as thoughtless as any other, work by simulating brain, or more generally biological, functions, as found for example in the field of biomimetic computation (e.g., Meyer, 1997). (The reasonable component of AI, in other words, would be better named SI, for either Substitute or Simulated Intelligence.) On the other hand, the study of intelligence-as-it-could-be, or life-as-it-could-be — which are supposed to be the subject matters of strong AI and AL (Langton, 1989) — have about as much chance of shedding light on biology, as the study of electricity-as-it-could-be, friction-as-it-could-be, or motion-as-it-could-be, have of illuminating physics.

5. Substance and Function II

Now another question: why is it that we were able to propose a stuff-general (not stuff-free) definition of memory (and perhaps of some other mental processes, such as computation), and an almost stuff-general definition of digestion, but not any definition of thinking apart from an activity of the brain, and likewise only a rather stuff-specific one of living systems?

The very concept of mental memory is a specialization of the overall concept of mentation, which could only be achieved because of the existence of the type of general understanding that our definition encapsulates. Indeed, for thousands of years or more, we have been using memory aides, or devices which help remember for us by taking over the memory function. In this way we have been able to discern memory as a task within a larger cognitive context, and so seek various ways of accomplishing it. And as luck would have it, memory as so defined is a ubiquitous function, of which mental memory is but one exemplar, and this was grasped and taken advantage of long ago. Likewise with computation, which in at least some of its aspects may not even be a native mental function, but one first acquired by abstracting from natural processes of accumulation and removal (for a view that at

least one portion of mental computation, namely the ability to do some counting, is native, see Butterworth, 1999). On the other hand digestion can be seen as a role in a larger metabolic process, this role, in our simplified characterization, being to supply energy. As luck would have it this energy must be found and stored in chemical bonds, as opposed to for example flywheels or atomic nuclei; and this is the reason why the definition is only 'almost' stuff-general. Digestion is also straightforward because it can be analyzed at the level of understandable component systems. Not even one entire ill-understood neuron can think, but one molecule can be digested.

By the same token, contrast memory with thinking in general: if we want to create a thing that is itself thinking, and not merely taking over the role of thinking within a larger process, then we must deal with thinking as an activity in and of itself — again, there is no difficulty in getting a non-thinking thing, like an abacus, to accomplish at least some tasks that we might accomplish instead by thinking. So, while there is nothing inherently wrong with definitions in terms of functional roles, i.e., in terms of effects instead of activities, such definitions have no bearing on the overall problems of mind or life, or more generally, autonomy of any kind. Consider then the various thinking aids, the ones we use to help accomplish cognitive tasks, such as language, diagrams, logic, mathematical symbols and rules, and abacuses plain and fancy. What do they have in common? They allow us to connect and manipulate various ideas by using symbols or tokens to represent them and their relationships. So thinking presumably has something to do with organizing and relating ideas. But we have no analogues of ideas to help us generalize any further. And we have no aides at all to help us construct radically new ideas, pose new problems, or evaluate which problems are worth solving and which are not. In other words, while we know a few non-mental things and processes that can be used to take over some of the more mechanical sub-tasks of thought, if we are not dualists or idealists, then we have no larger cognitive context within which thinking itself could be viewed as a task, and not simply an activity in and of itself. As a consequence, if we are not to use the term in an *ad hoc* manner, thinking can only be defined in terms of things that the brain does, regardless of whether or not any or even all of them can be prosthetically fulfilled. Similarly, if we want to create a thing that is itself alive, and not merely taking over the role of a living being in some specific system, then we must deal with living as an activity in and of itself. Consequently, if we are not to use the term in an *ad hoc* manner, life can only be defined in terms of things that organisms do.

It is not that we might not one day be able to propose a definition of, for example, the concept of intelligence — and why not simply the following: the joint ability to think new ideas and pose new problems, evaluate which ideas are worth entertaining and which problems worth solving, and to understand ideas and to solve problems. Nor is it that we might not one day be able to figure out what the brain does to accomplish such feats, or be, if we are lucky, able to prosthetically fulfil some functional sub-tasks. It is though that such knowledge will come in the

traditional manner of science, namely from studying the thing that does the doing. For example, to deepen this definition of intelligence, we must in turn define and explain the concepts of idea, understanding, problem, solution, and value; and the first two of these will have eventually to be defined in neurophysiological terms, such as neuronal assemblage, dendritic growth, lateral neuronal inhibition and the like. This is unaffected by whether or not the fantasy of a synthetic neuron is ever fulfilled.

We would do well to recall once more the situation with digestion, as either an activity or as a role within a larger metabolic process. While the definition we proposed in Section 3 may seem obvious to us today, we should remember that it would have been impossible to propose before substantial physiological investigation, and impossible to formulate before the substantial development of the underlying science of chemistry, involving as it does trans-phenomenal processes, and such high-level concepts as molecular bonds and chemical energy — even though a concept of digestion predates both sciences by millennia. Had hypothetical prehistoric workers in the hypothetically controversial field of Artificial Digestion, using advanced abacus theory (namely Turing's) without chemistry or physiology, been able to see into the future, they might have wanted to bear these facts in mind.

6. Substance and Function III: The Distinction Between the Two

Cognitivism is understood by many of its proponents to be the materialist answer to anti-scientific, or at least scientifically problematic, idealist and dualist psychology. Materialism is of course the core of the metaphysics of science, and cognitivists want very much to be scientific. The fundamental thesis of materialism is that the only real existents are material *things* (there are of course also fictional existents such as the characters in stories). Whether ordinary like rocks or exotic like photons, they are the only entities. Things have properties, like mass or energy or brightness, things have states (the current values of all their properties), and things undergo processes, like growth or decay or solidification, but, however precisely identified (for one proposal see Bunge, 1977, 1979; Mahner and Bunge, 1997), there are only things. There are no properties or states or processes or changes on their own. In particular, whether or not it is the only thing that may ever do so, it must be the material brain that thinks, and not the mind.

The cognitivist proposal to define mental states in terms of behavioural items, other mental states, and the causal relations between them, is therefore invalid. Materialism knows only states of things, and thus 'mental state' must be, as we use it, an abbreviation for 'a state, somehow special, of some material thing'. Nor do states actually do anything; instead it is the *actions* of one material thing, i.e., events, that cause changes in another, by means of a transfer of energy (Mahner and Bunge, 1997, p. 37f). Nor are all brain processes causal: neurons have spontaneous activities. And, while one might define any one mental state in terms involving a

different one, one cannot define all mental states, and the concept of mental state itself, in this way without circularity and uselessness: we want to know, finally, what a mental state is, not just how the various particular mental states are related to other particular mental states.

The role of materialism in science is not supposed to be merely decorative. It is supposed to be fundamental. That is to say, it, along with epistemological realism, methodological skepticism, and the principle of ontological lawfulness, is supposed to guide all analysis, help frame all conjectures, and control all arguments, keeping them grounded in reality.

Despite their professed materialism, cognitivists rarely employ it in their arguments. If they were to, they would not talk of causally interacting states or properties instead of causally interacting things, nor rule-governed processes instead of rule-fitting ones. (If rules could literally govern, i.e., have effects, then they would be sources or sinks of energy. And if processes could be governed instead of or in addition to things, i.e., if there were a process of controlling a process instead of or in addition to the process of controlling a thing, we would further have to contend with meta-meta-processes ad infinitum, and worry if any of them might be unstable, perhaps victims of feedback loops.) Sometimes metaphysical misfits like these may in context be innocuous. Often they are not: consider the following two.

In Section 3 we mentioned that cognitivists had, in addition to Rapaport's rebuttal of Dretske, another reason for disregarding Harnad's observation that a simulated fire is not hot. McMullin's (1997) reply to Searle's version of Dretske's premise, this time that simulated rainstorms do not make one wet, presents the objection as follows:

> ... Searle's analogy only begins to make sense if we already accept that minds are entities like rainstorms, whose realization demands certain specific, physical, causal powers, and are *not* entities like computers (or, if you prefer, computations), which can be realized by more or less arbitrary physical systems ...

This is an excellent illustration of what results when the fundamental thesis of materialism is not put to work during an argument's contruction. An appropriate reconstruction of it is:

> Searle's analogy only begins to make sense if we already accept that brains are entities like rainclouds, which are made out of only certain specific components, and are *not* entities like computers, which can be made out of more or less arbitrary components.

McMullin's formulation is a dead end: it suggests and indeed condones no line of investigation other than trial and failure. The materialist version, to the contrary and as always, invites both explanation and further inquiry: to begin with, recognize and distinguish between thinking as an activity of the brain (just as raining is an activity of rainclouds), and computers as devices designed and built to accomplish tasks. Then, study brains, or rainclouds for that matter, and find out exactly what if anything they do that computers of specific physical kinds (electronic, mechanical,

whatever) either do or do not. Instead of a mystery, one has a viable research programme. Sloman (1996) discusses some results of this inquiry that are of particular interest to computer scientists. For example, he shows how the theory of Turing machines is inadequate to model brains (see also Bunge, 1980, 1985).

The misidentification of substance and function, thing and property, entity and activity, is so pervasive in AI that even in major forums, the most elementary ideas from physics can be gotten disastrously wrong and overlooked. Consider the famous *Scientific American* debate between Searle and the Churchlands. There the latter (1990) proposed the following, of which only the last two or three claims were supposed to be false:

1. Electricity and magnetism are forces.
2. The essential property of light is luminance.
3. Forces by themselves are neither constitutive of nor sufficient for luminance.
4. Therefore electricity and magnetism are neither constitutive of nor sufficient for light.

This was meant to be analagous to Searle's Chinese Room argument (1980, 1990), with the goal to show that Searle's conclusion, namely that symbol manipulation is neither constitutive of nor sufficient for understanding, was just as false as the conclusion that electromagnetic radiation is neither constitutive of nor sufficient for light.

The comparison might have been intriguing, and so the validity of the analogy worth exploring, save for the small fact that the 'electricity' and 'magnetism' of the theory of electromagnetic radiation are not forces, they are *fields*. Electric and magnetic forces arise only when a suitable object interacts with such fields. Worse, there are no 'forces by themselves': all forces act on something. Thus in addition to proposition 1 being false, proposition 3 is both factually and vacuously true.

Then what if we place an emended proposition 3, saying that fields by themselves are neither constitutive of nor sufficient for luminance, in the role of the original? The Churchlands ask us to consider its hypothetical sway, had it been put forward by a naysayer like Searle at the birth of Maxwell's theory. They suspect it might have been compelling, in particular if buttressed by an experiment wherein a person lifted a magnet up and down without producing illumination. But they have got the historical context wrong too. First of all, "in the climate of understanding here contemplated — the 1860s —" no one would have paid the slightest attention to any experimentation so oblivious to the known facts of the day: long before the birth of Maxwell, let alone his theory, Herschel had already proved the existence of authentic invisible light of low frequency. Further, regardless of the existence of infrared light, or of the ultraviolet light discovered by Ritter one year later, and again contrary to the claim of the Churchlands, it was also already known long beforehand that whatever light was, it had only a certain range of frequencies, and that this range was in the neighbourhood of 5×10^{14} Hz, a trillion or so times higher than the frequency of the fastest arm movements. That is to say, unlike the

impression left by the Churchlands, it was already known long beforehand that light was a wave, and that whatever those waves were composed of, they could not be generated by someone waving their arms up and down in order to get the substance to oscillate.

When P.M. Churchland alone revisited the same argument in the *Journal of Philosophy* six years later, he edited out all reference to the historical context (apart from the suggestive illustration retained from *Scientific American*), although without acknowledging the erroneous presentation of it in the original. Instead of the climate of understanding of the 1860s, Churchland referred to the 'electromagnetically uninformed' — without we suppose intending to include those who believe light to be composed of forces, for this error persists explicitly in the more recent version. Nevertheless there is still a more basic item of knowledge that must be added to correct the setting.

It has been understood since prehistoric times that in order to produce a phenomenon, one requires not only the substances, but also that they be in the right states, or follow the right progression of states. For example, *Homo erectus* knew that wood and air are not flame, but that dry wood heated sufficiently, with enough air, burns. Similarly, arbitrary electric and magnetic fields do not luminance make: they have to oscillate as electromagnetic waves of the correct frequencies. Thus even the field-emended electromagnetic axiom 3 is still not precisely false. The Churchlands do not have the option of being precise, and pretending that electromagnetic fields, in the states specified by Maxwell, are lacking something light has; for their proposed experimental justification flouts Maxwell's conditions. This line of criticism does not apply to Searle, for in accordance with the view he criticizes, he specifies that the symbols be manipulated in whatever manner and at whatever pace necessary to produce the required output. That is why the standard AI objection to Searle's argument, one the Churchlands reject, is that the Chinese Room really does understand Chinese. Finally, while it is true that at the time of Maxwell, and indeed now, the Churchlands' original proposition 3 would have been tempting on its own, this is only because (a) unlike for fields, some background knowledge concerning forces is universal; and (b), consistent with that background knowledge, the proposition is in fact correct.

Rather than hypothetical history, why not consider what really happened: consider why Maxwell's proposal was taken seriously, investigated, and eventually accepted. Its foundations were a solid body of theory and experiment; the reasons why light should be identified with electromagnetic radiation, and the mutual mechanism of their action, were explained; it made for precise, testable predictions; and it produced mountains of profound results. Correspondingly, the identification of thought with brain processes is based on a solid body of evidence long accumulating; the mechanisms whereby the brain thinks are gradually being explained; by that explanation, we are gradually coming to a new and deeper understanding of the nature of thought; it makes for clear predictions, many of which have been tested and dramatically proved correct; and it is producing real and valuable results in

psychiatry, psychology, ethology, neuropharmacology and neurosurgery. Contrast this with the proposed identification of thinking, or more generally any process other than computation, with computation: it is based on no background evidence; in fact no one is really sure what the claim asserts; the original and only hypothesis of a mechanism whereby computers would think, namely by supposedly being just like the brain, a mere system of electrical relays, is known to be false; it has contributed no decisive predictions, if any correct ones at all; and the distinctive, valuable and true contributions it has made to the sciences of the mind, or more generally any science, are nowhere to be found.

7. The Turing Test

A word must be said about a famous alternative to our point of view, namely that of the Turing test (Turing, 1950; for subsequent proposals to make the Turing test more realistic, albeit still from a behaviourist or quasi-behaviourist perspective, see Harnard, 1991 and Schweizer, 1998; for additional arguments against the Turing test, see for example Block, 1995). In its essence, it says that the way to determine whether an entity is intelligent or not is to shield it from view, and then subject it to interrogation by an intelligent inquisitor. If by the responses the inquisitor cannot distinguish the candidate from a typical member of the inquisitor's own species, the Turing test is passed and intelligence certified. Note that this test is of no help in understanding or characterizing the basic nature of intelligence, and so of none *ab initio*: it proposes how to determine whether a second entity is intelligent, but not a first. Furthermore, it is the opposite of the scientific method, wherein objects are investigated not by obscuring them from view and artificially restricting the inquiry, but by taking every opportunity to delve into their inner workings.

If orthogonal to science, some say that the Turing test is instead of the common sense sort, based on our own daily experience and the way in which we determine that other human beings are indeed thinking beings. Surely such comments must only be in jest. For if you accept the premise of the Turing test in your daily life and are neither a higher vertebrate nor a sapient nor even a carbon chauvinist, you cannot take for granted that the other people (and dogs, cats and birds) you see on the street daily really do have mental lives, and must instead engage each one in a tricky and subtle process of questioning. Nor on the other hand can you take for granted that such things as microbes, computers, and omelettes do not have mental lives, and must also question each bacterium individuallybefore boiling water, each computer before pulling the plug, and each breakfast before eating it. (Or at the very least, a first exemplar of each kind. But a true functionalist should not take for granted that just because a first bacterium or computer or breakfast might be found dumb, so too will all the others; after all, if the same or similar functions can be served by very different things, why should not the same or similar things serve very different functions now and again?)

But suppose for a moment, restricting ourselves to circumstances closer to those for which the Turing test was originally envisioned, that you really did one day create a thinking, English-speaking computer. But suppose in addition you had succeeded too well, so that not only was your computer sentient, but also shy. Suppose it was so shy that it tended to choke under pressure. Surely there could be no more pressurized situation for a shy computer than to have to undergo the Turing test in defence of its very life. Confronted thus, the shy computer would indeed choke and produce no output; and here again the Turing test would be of no help at all, failing us when we needed it the most.

Turing acknowledged other objections leading to similar conclusions, and allowed that while passing his test was not a necessary criterion for intelligence, surely it was a sufficient one. Yet he could not answer one simple question: why? Why should it take intelligence to convince a man to affirm it? Anyone who has squared himself with human history would not have had to wait for the first public attempt at a formal Turing test[5] to prove decisively that it does not. The members of *Homo sapiens*, Intelligent Man, have frittered away the lives of countless generations praying to wooden idols and stone statues, spilled the blood of untold thousands if not millions on sacrificial altars to nothing, and waged war after ruinous war under orders from mute gods. The truth of human history is that it requires neither intelligent behaviour nor behaviour at all, nor even existence, to convince countless men no less intelligent than any other, and even a few Nobel Prize winners, of the presence of not just intelligence, but supra-intelligence.

8. Some Variants of Anti-Functionalism

Functionalism has been well and justly criticized by several philosophers (e.g., Block, 1980b; Bunge, 1985; Searle, 1992; Schweizer, 1996). While we support Searle's anti-functionalism, we are not precisely in agreement with his metaphysical and epistemological alternatives, or his characterizations of materialism, monism and objectivity. For example, we disagree with his view that brains *cause* consciousness, because we accept only efficient causation, i.e., causation as energy transfer, and for brains to cause consciousness would then require that brains transfer energy to some other entity thereby made conscious. In other words, the relation of emergence, which, as we agree with Searle, is at issue here (see also Bunge, 1979, 1980; Humphreys, 1997), is in our view not a causal relation. Just as things do not cause their basic properties, as bodies do not cause their mass, so too the emergent properties of systems, such as the consciousness of some brains, are not caused by them, but simply inhere in them. (For a microencapsulated overview of our basic ontological viewpoint, see Table 1.)

Likewise, although we agree with Harnad (1991, 1993) on the distinction between original and simulation, we disagree with his views on the possibility of understanding the nature of the mental, and with his contention that there can be no evidence for mental capacities other than our own. Harnad claims that mental

Table I. Microencapsulated overview of our basic ontological viewpoint (for details see Bunge, 1977, 1979; Mahner and Bunge, 1997; and references therein)

Tenet	Meaning	Confused with	Example or explanation
Reality is objective (ontological realism)	World exists on its own, without help of observers	Impartiality; world accessible to inquiry; intersubjectivity; there are no subjective experiences	Contrary view violates scientific laws: e.g., discovery of Uranus would then have violated conservation of mass and energy
All existents are material (materialism)	There is only one fundamental substance, matter (as opposed to, e.g., mind or spirit); its fundamental property is energy (capacity for change); there are no processes apart from things	Everything is a classical particle; everything is solid; everything has mass; there are no thoughts or feelings or processes of any kind	Everything is changeable and can interact with other things, thus constituting one ontological realm, that of matter; processes are successions of changes *of* things, not new things on their own
Systemism and emergentism	Things may conjoin and result in systems, which have properties their components lack; likewise, components have properties their systems lack	Emergent properties are those that cannot be explained in terms of the properties of components; holism; mysticism; supervenience; devaluation of the individual	No single water molecule is a liquid, but large systems of water molecules can be; normally functioning persons have a sense of smell, but societies do not
Ontological lawfulness, incorporating chance and accident	All processes, even stochastic ones, are lawful; accidental circumstances can arise from the conjunction of independent trajectories	All processes are algorithms; there is no free will; reality is oppressive	All things have their own nature, which may include stochastic behaviour; regularities emerge from large numbers of accidents
Psychological materialism (monism of the materialist kind)	Thoughts are brain processes	Thoughts are caused by brain processes; the brain 'secretes' thoughts; computers think; thoughts are algorithms; there are no thoughts	Work of Penfield, Hebb, their predecessors and successors

facts cannot be inferred via the scientific method of conjecture, argument, observation and experiment in the same way that any other facts can, since no one could ever tell the difference between an entity with mental capabilities, and one without them but which always behaved exactly as one with. Harnad argues that there is no room for such difference without eventual observable consequence in physics, or even biology, but that there is in psychology. The distinction lies supposedly in the reality of the laws of physics, and presumably, biology, which we are led to believe cannot be the case for psychology.

But why should the mental be any different from anything else in this regard? Fundamental to the scientific world view is the principle that *all* properties[4] are lawful, i.e., related to other properties in an invariant way. This holds for the mental just as much as for anything else, and for the same reasons. We submit that no one could possibly even imagine a truly lawless external world, where all was erratic, the future unlinked to the present, and the present not having come forth from the past, for in it every thing would change to an entirely different thing, or nothing at all, at every instant. For exactly the same reasons, we submit that no one could possibly even imagine what it would be like for our mental properties to be lawless either: one's thoughts at one instant bearing no relation to one's thoughts at the next; the impossibility therefore of having a train of thought; no chance to concentrate or ponder or imagine; and of course no possibility of memory, as our original characterization makes explicit.

Indeed, the principle of lawfulness underlies all of our own anti-functionalist arguments, and our assertion that mental capacities are of brains and nothing else. The special properties of complex systems, such as the mentation of brains, must emerge lawfully out of the properties of the system components and environment, and from the way they are all linked together. Thus in the brain, mentation emerges from the uncounted special properties of the neurons, synapses, et alia; the influence of sensory input, and the overall pattern of neural organization. Since properties are bound lawfully, one is simply not free to create any and all higher-level features of reality out of just anything.

Likewise, it is a basic methodological adjunct of the principles of lawfulness and materialism that, if things are somehow different — as a mindless zombie knock-off is hypothesized to be different from the mindful original — then those differences can be probed, at least in principle and if only indirectly: since all properties are lawful, and all things can in principle interact with other things, i.e., have their properties altered by that interaction, then from the alterations in the observable properties, we can infer via conjecture, argument, and further observation and experiment to the unobservable ones.

And if it should happen that this most fundamental ontological and epistemological principle of science, namely that all unobservable properties are somehow eventually related to observable ones, is false, then again we are in no worse shape with the mental than with anything else. For the possibility of some subjectively knowable aspect of our mental lives being inscrutable in principle to science is

no greater than the possibility of protons or magnets having secret lives of their own, forever inaccessible to inquiry. And in all seriousness, while we cannot immediately discard the latter possibility, we can the former, for we do know that the mental facts typically referred to as only subjectively accessible, *are* connected to objectively observable ones, since we ourselves make them accessible to others all the time, by our actions voluntary and involuntary. And the causal chains go inwards too: every last aspect of our own mental lives that we ourselves subjectively know about can be impinged upon from the external world. This is the basis of both the investigations and the real results of the modern science of psychology. For example, a modern psychologist will not simply take free will at face value, let alone ignore it, but investigate how it is constrained, shaped, made stronger or ruined by such things as training, drugs, fear, blows to the head, neurosurgery, propaganda and nutrition.

9. Conclusion

Although most functionalists consider themselves materialists, functionalism is in our view basically Platonist, for maintaining that form and function do not inhere in matter, but are just incidentally carried by it. Materialists hold on the contrary that life cannot be studied apart from the matter that does the living, namely organisms, nor mind from brain. Thus materialists are not surprised to discover that the biological literature is a vast repository of knowledge about life, while the (strong) artificial life literature is not; nor that the neuroscientific approach to psychology (e.g., Hebb, 1949; Bunge, 1980; Bunge and Ardila, 1987; Kosslyn and Koenig, 1995; Mountcastle, 1998), as recent as it is, has similarly trounced the ones and zeros approach, which, for all the time, money and effort poured into it, is for example yet to result in a valid therapy, whether for the psychiatry of humans or machines.

What then has been the contribution to either biology — the science of life — or psychology — the science of mind — of the functionalist and computationalist programmes of strong AI and AL? In our assessment, they have served mainly to co-opt good brainpower to the video game and speculative metaphysics industries. This judgement might be considered harsh, and even unfair, were it not for the numerous workers in the AI and AL fields who would take it not as condemnation but as both fact and point of pride, industrial epithet aside. They see no intellectual reason why anyone who claims to be studying or producing artificial intelligence or artificial life must be held to a standard requiring them to produce or use respectively knowledge about natural intelligence and natural life, or else be found to be using the key words arbitrarily. No wonder they claim that in order to understand AI or AL, we must think of life and mind in an entirely new — translation, *ad hoc* — manner. Surely there are many others who believe that there is a *tertium quid* between materialism and functionalism, but they are not able to articulate a metaphysics that could accommodate these two mutually contradictory world

views. Nor have they been able to offer a positive reason why anyone would want to; they have only the compelling negative reason that otherwise, AI and AL will be indistinguishable from science fiction (Pattee, 1995). And these are still fields where the publishing of a new video game can be lauded as a seminal event in creation, to the extent that the authors or programmers consider themselves to have thereby attained godhood (Grand, 2000).

It is true that many neuroscientists resort uncritically to popular computerspeak and infospeak. It is also true that many serious biological and psychological investigations are being helped by the wholesale use of computation, and vice versa. Nevertheless our understanding of mental function is still based on endogenous neuroscientific and psychological concepts such as those of neuronal assemblage, neurotransmitter deficiency, dendritic growth, hormonal balance, lateral inhibition, affect, mood, disposition, learning and the like, not computer science concepts such as recursion, illegal instruction, bit shift, parity check, array, pointer, counter, stack, file allocation table, file transfer protocol, file format, root directory, passing of parameters, machine versus assembly language, compiled versus interpreted language, object-oriented programming and functional programming, NP-completeness, declaration of variables, or the like. Essentially all technical concepts from computer science which did not already have a pre-existing popular meaning, such as information, memory, or computation itself, are entirely alien to all modern knowledge of both life and mind. Where the science and technology of computation have contributed to biology and psychology, it has been through simulation and solution, not by the transplantation of any understanding. If we are to use the theory of computation to understand life and mind, it looks like we will have to understand computation in an entirely new — translation, *ad hoc* — translation, biological, manner. Equivalently: the natures of life and mind are material, not computational or more generally, formal.

We respond to our original question as follows: you will know you have synthesized a thinking thing when you have synthesized a working brain. This conclusion is not a triumph of conservatism but a consequence of the principles of ontological lawfulness, materialism, epistemological realism, and methodological skepticism; and finally of the results of science. The more we study the brain and living things, the more we realize how extraordinary and valuable they really are. The proponents of strong AI and AL are not as impressed: while none of them would even dream of building something as mundane as a television set out of curds and whey, for some reason there is no shortage of those who see no obstacle, in principle, to building a living, thinking, feeling being out of Swiss cheese.

Notes

[1] Putnam has since renounced this position (1994).

[2] As one anonymous reviewer dismissed our arguments: "one can see this first hand in my own lab (and thousands of labs like it). In this lab, robots are under construction. The idea is not that these robots are models of animals. The idea is rather that these robots *are* animals."

[3] By this and similar expressions, we mean an activity in a material system (the brain, in our view) that is of mental character, not, as some would take it to mean, a process occurring in an immaterial mind.
[4] We refer here only to what are known as essential properties, i.e., those that make a difference to the nature of the entity, in contrast with accidental properties.
[5] This was held at the Boston Computer Museum in 1991. Tested against J. Weizenbaum's simple ELIZA automatic psychobabble generator, five out of 10 judges determined that ELIZA was really human (Shieber, 1994).

References

Arkin, A. and Ross, J. (1994), 'Computational Functions in Biochemical Reaction Networks', *Biophysical Journal* 67, pp. 560–578.

Block, N. (1980a), 'What Is Functionalism?', in N. Block, ed., *Readings in Philosophy of Psychology*, vol. 1, Cambridge, MA: Harvard University Press, pp. 171–184.

Block, N. (1980b), 'Troubles with Functionalism', in N. Block, ed., *Readings in Philosophy of Psychology*, vol. 1, Cambridge, MA: Harvard University Press, pp. 268–305.

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*, Cambridge, MA: MIT Press.

Boden, M.A. (ed.) (1996), *The Philosophy of Artificial Life*, Oxford: Oxford University Press.

Boden, M.A. (1999), 'Is Metabolism Necessary?', *British Journal for the Philosophy of Science* 50, pp. 231–248.

Bunge, M.A. (1977), *Treatise on Basic Philosophy, vol. 3: Ontology I*, Dordrecht: D. Reidel.

Bunge, M.A. (1979), *Treatise on Basic Philosophy, vol. 4: Ontology II,* Dordrecht: D. Reidel.

Bunge, M.A. (1980), *The Mind-Body Problem*, Oxford: Pergamon Press.

Bunge, M.A. (1985), *Treatise on Basic Philosophy, vol. 7, Part II: Life Science, Social Science, and Technology*, Dordrecht: D. Reidel.

Bunge, M.A. and Ardila, R. (1987), *Philosophy of Psychology*. New York: Springer.

Buschlinger, W., Vollmer, G. & Walter, H. (1998), 'The Mental Life of Artifacts. Explications, Questions, Arguments', *Zeitschrift für Naturforschung* 53c, pp. 455–479.

Butterworth B. (1999), *What Counts: How Every Brain is Hardwired for Math*, New York: Free Press.

Chandrasekaran, B. and Josephson, J.R. (1997), 'Representing Function as Effect', in M. Modarres, ed., *Proceedings of the Fifth International Workshop on Functional Modeling of Complex Technical Systems*, College Park: Center for Technology Risk Studies, University of Maryland, pp. 3–16.

Churchland, P.M. (1996), 'The Rediscovery of Light', *Journal of Philosophy* 93, pp. 211–228.

Churchland P.M. & Churchland P.S. (1990), 'Could a Machine Think?', *Scientific American* 262(1), pp. 32–37.

Dennett, D.C. (1978), *Brainstorms: Philosophical Essays on Mind and Psychology*, Montgomery: Bradford Books.

Dretske, F. (1985), *What Computers Can't Do: The Limits of Artificial Intelligence*, New York: Harper & Row.

Grand, S. (2000), *Creation: Life and How to Make It*, London: Weidenfeld & Nicolson.

Harnad, S. (1991), 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1, pp. 43–54.

Harnad, S. (1993), 'Artificial Life: Synthetic Versus Virtual', in C.G. Langton, ed., *Artificial Life III*, Redwood City: Addison-Wesley.

Hebb, D.O. (1949), *The Organization of Behavior*, New York, Wiley.

Humphreys, P. (1997), 'Emergence, not Supervenience', *Philosophy of Science Supplement* 64: S337-S345.

Kary, M. (1990), 'Information Theory and the *Treatise*: Towards a New Understanding', in P. Weingartner and G.J.W. Dorn, eds., *Studies on Mario Bunge's* Treatise, Amsterdam: Rodopi, pp. 263–280.

Kosslyn, S.M. and Koenig, O. (1995), *Wet Mind: The New Cognitive Neuroscience*, New York: Free Press.

Langton, C.G. (1989), 'Artificial Life', in C.G. Langton, ed., *Artificial Life*, Redwood City: Addison-Wesley, pp. 1–47.

Langton, C.G. (1991), 'Introduction', in C.G. Langton, C. Taylor, J.D. Farmer and S. Rasmussen, eds., *Artificial Life II*, Redwood City: Addison-Wesley, pp. 3–23.

Mahner, M. and Bunge, M.A (1997), *Foundations of Biophilosophy*, Berlin, Heidelberg, New York: Springer.

Mahner, M. and Bunge, M.A. (2000), *Philosophische Grundlagen der Biologie*, Berlin, Heidelberg: Springer.

Mahner, M. and Bunge, M.A. (2001), 'Function and Functionalism: A Synthetic Perspective', *Philosophy of Science* 68, pp. 75–94.

McMullin, B. (1997), 'Computing Machines and Mentality', *Brain and Cognition* 34, pp. 28–47.

Meyer, J.A. (1997), 'From Natural to Artificial Life: Biomimetic Mechanisms in Animat Designs', *Robotics and Autonomous Systems* 22, pp. 3–21.

Mountcastle, V.B. (1998), *Perceptual Neuroscience: The Cerebral Cortex*, Cambridge, MA: Harvard University Press.

Pattee, H.H. (1989), 'Simulations, Realizations, and Theories of Life', in C. Langton, ed., *Artificial Life*, Redwood City, CA: Addison-Wesley, pp. 63–77.

Pattee, H.H. (1995), 'Artificial Life Needs a Real Epistemology', in F. Moran, A. Moreno, J.J. Merelo and P. Chacon, eds., *Advances in Artificial Life*, Berlin, Heidelberg: Springer, pp 23–38.

Putnam, H. (1975), *Philosophical Papers*, vol. 2, New York: Cambridge University Press.

Putnam, H. (1994), 'Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind', *Journal of Philosophy* 91, pp. 445–517.

Pylyshyn, Z.W. (1984), *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, MA: MIT Press.

Rapaport, W.J (1988), 'Syntactic Semantics: Foundations of Computer Network Language Understanding', in J. Fetzer, ed., *Aspects of Artificial Intelligence*, Dordrecht: Reidel, pp. 81–131.

Schweizer, P. (1996), 'Physicalism, Functionalism and Conscious Thought', *Minds and Machines* 6, pp. 61–87.

Schweizer, P. (1998), 'The Truly Total Turing Test', *Minds and Machines* 8, pp. 263–272.

Searle, J.R. (1980), 'Minds, Brains and Programs', *Behavioral and Brain Sciences* 3, pp. 417–424.

Searle, J.R. (1990), 'Is the Brain's Mind a Computer Program?', *Scientific American* 262(1), pp. 26–31.

Searle, J.R. (1992), *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Shieber, S.M. (1994), 'Lessons from a Restricted Turing Test', *Communications of the ACM* 37(6), pp. 70–78.

Sloman, A. (1996), 'Beyond Turing Equivalence', in P.J.R. Millican and A. Clark, eds., *Machines and Thought: The Legacy of Alan Turing* (vol 1), Oxford: Clarendon Press, pp. 179–219.

Sober, E. (1991), 'Learning from Functionalism — Prospects for Strong Artificial Life', in C.G. Langton, C. Taylor, J.D. Farmer and S. Rasmussen (eds.), *Artificial Life II*, Redwood City, CA: Addison-Wesley, pp. 749–765.

Turing, A.M. (1950), 'Computing Machinery and Intelligence', *Mind* 59, pp. 433–460.