Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 82 (2013) 12 – 18

Procedia
Social and Behavioral Sciences

World Conference on Psychology and Sociology 2012

Artificial Psychology, Functionalism and Mental Representation

Nazim Gokel [a] *

[a]*Philosophy Department, Boğaziçi University, Bebek, Istanbul 34342, Turkey*

Abstract

In this paper, I aim to reveal and examine the ideology behind artificial intelligence and functionalism from a philosophical perspective. To this aim, I first give a short review of functionalism. Then, I move on to explain the Platonic ideology, which, according to me, is deeply rooted in almost all versions of functionalism. I argue for the claim that the software/hardware dichotomy is only a pragmatic distinction and one should not expect to find mutually exclusive categories signified by software and hardware terms. In the final section, I suggest that there could be a better, more sensible, metaphysics at the background of functionalism, one that may offer a much more fruitful way to understand the nature of representation and mind.

Keywords: Artificial Psychology, Functionalism, Platonic Ideology, Intentionality, Representation.

1. Introduction

There is something extremely incredible and stunningly intuitive in the ideology behind artificial intelligence in computer science and functionalism in philosophy of mind. It is my initial aim, in this paper, to reveal and examine this ideology from a philosophical perspective. To this purpose, I first give a very short sketch of functionalism. Then, I move on to explain the Platonic ideology, which, in my view, is deeply rooted in almost all versions of functionalism. This ideology is often manifested with the "software"/"hardware" dichotomy. I argue for the claim that the software/hardware dichotomy is only a pragmatic distinction, and it should not be taken to imply metaphysically exclusive categories. Having shown the misconception about this distinction, then, I will suggest, in the last section, that there could be a better, more sensible, metaphysics at the background of functionalism, one that may offer a much more fruitful way to understand the nature of representation and mind.

* Corresponding author: Nazım Gokel.
E-mail address: ngokel@aol.com

doi:10.1016/j.sbspro.2013.06.218

2. Artificial psychology and functionalism

Artificial Intelligence (shortly as AI) is a research field that aims to investigate the question of whether it is logically and technically possible to build machines or program computers that can achieve diverse sorts of cognitive activities, activities that involve deductive and inductive reasoning, believing, desiring, planning, dreaming, etc., much like the activities of which normal adult human beings are capable. For Simon and Newell, building computers that can think was not only logically possible, they believed that soon we will be witnessing that a digital computer will beat the world's chess champion, it will be able to discover and demonstrate a new mathematical theorem, it will compose a fine piece of music, and most importantly with these advances in AI, "most theories in psychology will take the form of computer programs, or of qualitative statements about the characteristics of computer programs" (Simon & Newell, 1958). This could have been the most robust expression of the initial targets and ideology behind AI. Simon and Newell put forward a firm and strong claim that there is no contradiction in terms about the idea of artificial psychology. After one point, there will be probably no sense to distinguish artificially intelligent systems from naturally intelligent organisms. We will eventually have the same moral obligations toward them, at least so Simon and Newell might believe.

On the other hand, a few philosophers in the late-60's, who were deeply fascinated by the computer model of mind and Turing machines, began to develop a functionalist theory of mind (henceforward as FTM) that updates behaviorism and mind-brain identity theory (Block, 1980; Block, 1995; Kim, 1996). Since then FTM has been the cornerstone of the contemporary philosophy of mind. Functionalism achieves two things. First, as opposed to the behaviorist framework, it prominently demonstrates the fact that mental states must be inner states with functional-causal roles. Secondly, as opposed to chauvinistic attitude in the mind-brain identity theory, it opens up a widely liberal perspective through which any kind of being/system (e.g. computers, souls, extra-terrestrials) that can pass Turing Test is qualified as an intelligent being. In a way, FTM has brought forth a radical change to our way of describing mentality and intelligence by offering a new conceptual repertoire for the comprehension of the metaphysics of mind, a framework that was not only backed up by largely optimistic AI research predictions but also pumped up with extremely powerful forces of intuition and imagination.

In order to pinpoint the underlying ideology behind FTM, it is crucial to provide a brief sketch of the theory in a few lines beforehand. With this brief presentation, I hope, one will be able to come to recognize its thought-provoking and intuition-based reasoning features. Without further ado, let us then give an outline of Functionalism. Consider the following simple example of keys. A key can be realized by many diverse physical kinds. For instance, it can be made out of metal, plastic, wood, and so forth. It may also come in many different shapes, weighs, colors, etc. But, all these keys have one thing in common, that is to say that their job is to unlock something like door locks, bank vaults, etc. Without this unique role, none of them would be counted as a "key". By the same token, functionalists argue for the claim that the intrinsic/essential property of a mental state has to do with its causal role in a network of functional architecture of a being/system, and it has nothing to do with what realizes or occupies that role. Hence, there is the distinction between role and occupant of mental states. A mental state can be realized by physically, and even spiritually, different beings, and provided that a state has the same functional role across different physical/spiritual beings, then we are entitled to assign the same mental state to all these beings. It is even possible for the same individual to entertain the same mental state when the state in question is realized by different physical-chemical states of the individual at different times. More explicitly, FTM states the following claim:

Any being/system B is in a type M of mental state, if there is a state of B that plays a particular functional/causal role F (definitive of that M-type) within a complex network of states of B such that it mediates, together with other relevant states in that network of states, between perceptual inputs and behavioral outputs of B.

It is a purely empty and futile enterprise, according to the functionalist metaphysics of mind, to explore the source of psychological laws, which govern mental operations and behaviors of material or spiritual beings, in

the underlying (material/spiritual) structure/substance where those laws make their first appearance. This is another way of saying that there is no difficulty, in principle, for the individuation of the same mental states in physically/spiritually differently composed beings and substances, provided that those beings and substances manifest functionally equivalent architectures, i.e. a functional organization of an appropriate kind that is constitutive of mentality and intelligence. Viewed in this manner, functionalism surprisingly makes both materialists and spiritualists happy, probably the first time in the history of mankind. For instance, let us have a look at the functionalist description of pain for material beings:

A material subject, M, is in pain, if there is a functional state, say F10, such that it mediates, together with the other functional states, between bodily damage and avoidance/escape behavior from the source of that bodily damage.

And now have a look at spiritual pain:

A spiritual subject, S, is in pain, if there is a functional state, say F10, such that it mediates, together with the other functional states, between spiritual damage (as the effect of psychological disturbance of another soul) and avoidance/escape behavior from the source of that spiritual damage. [Both examples belong to me]

Any mental state, intentional or phenomenal, can be realized in material or spiritual substances, and it just does not matter at all in which kind of substance it is *embodied* or *ensouled*. The material or spiritual host of mental states does not contribute to the essence of those states; they are required only for the realization of those states. In a sense, they help them come into being, but they do not determine the essence of mental states. But, this gives rise to a number of questions: Can functionalists successfully explain the essence of mental states merely in terms of functional roles? What does "functional organization of appropriate kind and complexity" amount to exactly?

To begin with, mental states are not only inner states with functional-causal roles; they also have intentional and qualitative properties. For simplicity, let us take a belief as the paradigm example of mental states. A belief, from a functionalist point of view, is a monadic functional state, say F10, with a unique causal role such that, whenever such-and-such input conditions obtain, it typically leads to another functional state(s), say F11, which together typically terminate in such-and-such output patterns. At this point, it is not very clear, however, whether functionalism can reduce or explain intentional and phenomenal characteristics of mental states purely in terms of functional-causal roles. According to Brentano, one of the touchstones of mentality is that mental states represent things in the world as being a certain way, that is to say they are all about something in particular. Moreover, intentional systems are capable of entertaining thoughts that involve things not present/available at the moment of thinking or, even, things that are non-existent objects in the world. These two features are commonly called "aboutness/directedness" and "intentional in-existence". Furthermore, a subject's belief may accompany a qualitative experience, what it is like to be in that state. Qualia may be, for some people, one of the most important ingredients of mentality without which we may never be said to be conscious of anything.

Especially with the impact of Putnam's Twin-Earth thought experiment (Putnam, 1975), Searle's Chinese Room argument (Searle, 1980) and Block's China Nation thought-experiment (Block, 1978), a great number of people began to suspect (1) whether it is really true that if A and B are functionally equivalent, then A and B cannot differ with respect to contents/meanings of their functional states (Shagrir, 2005), (2) whether functional duplicates can instantiate authentic intentionality, (3) whether functional duplicates are capable of thoughts and consciousness. Let this suffice for a short review of functionalism. In the sequel, I will attempt to highlight the underlying ideology behind FTM.

3. The Platonic ideology

I tend to believe that there exists a hidden Platonic ideology behind functionalism, operative in the background of almost each version of it. This ideology, a two-leveled explanatory framework, is often manifested with its constant emphasis on the software/hardware dichotomy. A few words about Plato's metaphysics are in order.

In Plato's metaphysics, Ideas/Forms stand for the perfect, eternal, never-changing paradigms or archetypes that exist in the realm of Being; whereas objects/things in the realm of Becoming are imperfect and subject to change over time. Platonic Forms are, in a sense, ideal, (ultimately) real and perfect models in which particular things in the realm of Becoming share or participate. In other words, those ordinary objects are only exemplifications or realizations of Forms. Logically speaking, the very existence of ordinary objects (logically and ontologically) depends on the existence of Forms, because without any Form that exists to be shared, there could not be ordinary objects too (Watson, 1995). Thus, the Form X exists independently of X-ing objects, but X-ing objects cannot exist without the Form X. Metaphorically speaking, objects in the material world are like shadows of those ultimately real Forms in the world of Being. Within this context, a soul, according to Plato, is a miserable and unhappy prisoner in a body with forgotten memories, and its freedom and capacity to know the real nature of things lies in its own intellectual revolution by means of which it comes to a gradual recognition of the fact that ultimate reality lies beyond the shadows of the material world.

It is striking to notice that FTM is deeply rooted in a Platonic metaphysics. In much the same spirit, FTM asserts the claim that the function of something (e.g. keys, coke-machines, computers) is independent of the physical aspects of what realizes it. Then, it follows, according to functionalists, that a mental state must be identified with a functional state that is independent of the physical/spiritual make-up of whatever *embodies* or *ensouls* it. In functionalism, there is a two-leveled conceptual scheme according to which any given mental state is described with reference to its physical/spiritual composition or its (potential or actual) causal role in a vast array of functional architecture. From a functionalist point of view, lower-level aspects of mental states (e.g. neural firings, synapses in the neural networks) only indicate what mental states are actually made of with respect to one kind of intelligent beings (e.g. human beings and some animals), but they do not explain what constitutes/makes up mental states in general. A pain, for instance, can be realized by C-fiber activations in human beings, or it can be realized by D-fiber activations in octopuses, or, even, it can be realized by some kind of plasma in extra-terrestrial creatures. There must be a universal pain in virtue of which (physically/spiritually) different kinds of pain-suffering beings/systems are said to be in pain. To explain this conceivability of a universal psychology, FTM provides a very liberal account of mind according to which mental states are, in essence, functional states. Functional states, in turn, stand for higher-level properties of intelligent beings, intrinsic properties that give something a definitive character, e.g. being a kind of intelligent being/system. For FTM, it is not the fine structural details about the material/spiritual host, but the functional details about the complex causal network that determine whether a being/system has a share in universal mentality and intelligence. Functionalism, therefore, rests on the assumption that there is a universal psychology that can be shared across physically/spiritually diverse beings/systems. This ideology is often illustrated with the software/hardware dichotomy. Heil (1998) says that:

> Every program is "embodied," perhaps, in some material device or other. But the very same program can run on very different sorts of material device. In the same vein, we might suppose that every mind has some material embodiment, although minds may have very different kinds of material embodiment. In the case of human beings, our brains constitute the hardware on which our mental software runs. Alpha Centaurians, in contrast, might share our psychology, our mental software, yet have very different, perhaps non-carbon-based, hardware. (p. 91)

This is one of the fine instances of the computer model of mind according to which mind is nothing but a universal software that can run on radically different sorts of physical/spiritual stuff. Intelligent beings are intelligent, because all of them share the same universal psychology/software regardless of the differing (material or spiritual) ways they realize that psychology/software. This way of characterizing mind, however, gives rise to a number of questions. As far as I can see, there could be two groups of questions, the former of which is related to the identity conditions of a software and the latter of which is related to the emergent need to clarify notions of "software", "hardware" and "computer". So, let us begin with the first group of questions.

Can there really be a universal software/program that can run on different kinds of hardware? The very idea of universal software implies the claim that it is possible to run exactly the same program in physically diverse machines/hardware. But, the following question must be answered first: Under which conditions it is true to say that X is exactly the same program running on different machines? From my point of view, the idea of universal software must be, at least, a matter of debate, yet most people find it obvious and straightforward. Consider the following case. Suppose that you have a program running so smoothly on one of the most advanced computer hardware technologies available now, a computer that is composed of high-tech silicon chips and electronic circuits. Just think what would happen if you want to use exactly the same program in old computers composed of valves and gears. Can it be installed in the first place? One can plausibly argue for the claim that the way to program old computers must be extremely different from the way to program recent computers, because their computational architecture is radically different due to the physical structure of their components. Installing the same program, however, might not be the real problem after all; one can find a good medium to do the work of installation. But, one can go on to insist that installing the same program will require much labor to modify the algorithm in order to adapt computational and physical capacities of the old hardware. So, if the algorithm is the heart of any software, then changing the structure of an algorithm is nothing but playing with the identity conditions of software. If this is true, then we can no longer talk about the identity of software over physically diverse machines (see Eliasmith, 2002).

As for the second set of questions, it is not really clear whether software is something entirely independent of hardware. In general, functionalists seem to hold a view that software-properties of X mark a completely different and independent category than those hardware-properties of X. Does this distinction between software-properties of X and hardware-properties of X really entail that software is entirely independent of hardware? Or, broadly speaking, does the distinction between pure functional properties and pure structural/physical properties entail that function is independent of matter? Let us have a look at Moor's warning about the misconception of "software" and "hardware". Moor (1978) points out that:

> Computer programs are often considered to be part of the software of a computer as opposed to the hardware. Unfortunately, computer hardware is frequently characterised as 'the physical units making up a computer system' (Chandor [1970], p. 179). By contrast this seems to suggest that software is not part of the computer system or even worse that it is not physical. It is important to remember that computer programs can be understood on the physical level as well as the symbolic level. The programming of early digital computers was commonly done by plugging in wires and throwing switches. Some analogue computers are still programmed in this way. The resulting programs are clearly as physical and as much a part of the computer system as any other part. Today digital machines usually store a program internally to speed up the execution of the program. A program in such a form is certainly physical and part of the computer system. Furthermore, since programming can occur on many levels, it is useful to understand the software/hardware dichotomy as a pragmatic distinction. (p. 215)

The whole point boils down to the issue whether software-properties and hardware-properties of a computer indicate a metaphysical distinction or a pragmatic distinction. Moor has a way to settle this issue. For him, one can describe any given computer in terms of its physical aspects or symbolic aspects. For instance, a computer, in a physical description, is an entity composed of silicon chips, electronic circuits and so forth. If you hold on to this description and know enough about the laws of physics about electronics, Moore suggests, you will be able to predict any computer behavior. In addition, one can describe any given computer with respect to its symbolic features. Consider a chess-playing computer. In this case, a computer is a system that can take certain kinds of inputs as symbols, go to a set of states relevant to that input, make use of its database and algorithms, and choose the best output in accordance with inputs. If you detect, for instance, that the computer always moves its queen too early, and know enough about the logical structure of the computer, then you may easily spot the source of the problem (Dennett's original example is in Dennett, 1978, p. 107; see also Cummins, 1986; Skokowski, 1994).

By following this strategy, for Moor, one can also provide two different ways to characterize a computer program, the former of which might emphasize its physical properties such as "series of punched cards, configurations on a magnetic tape", while the latter might take its symbolic properties such as "a set of symbolic instructions" (Moor, 1978, p. 213). In the final analysis, Moor wanted to show that software and hardware terms do not actually signify mutually exclusive categories. Just because one can explain software properties of a computer without a reference to its physical aspects, this does not, in any sensible way, prove that software properties are independent from hardware properties. The software/hardware distinction is a pragmatic distinction, and programming, as Moor and Lycan made painstaking efforts to show in detail, is relative to one's perspective (Lycan, 1990; Moor, 1978).

4. Mental representation:

Instead of explaining Form as something connected with Matter, Platonic metaphysics takes it to have a higher-order mode of being, one that is independent of ordinary material objects and the logical/ontological ground of them. Similarly, FTM takes functional properties of mental states as higher-order properties that are independent of physical/spiritual properties of their realizers/hosts and the existence of those functional properties in any body/soul actually explains the reason of that body/soul being intelligent and conscious. In particular, FTM accepts a very Platonic idea that there is universal software that can run on and be shared across physically/spiritually diverse kinds of beings/systems. At the end of the previous section, I provided a number of reasons why this way of characterizing mentality and intelligence is very problematic. Alternatively, one can adopt an Aristotelian metaphysics according to which the Form of something is intimately connected with the Matter of that being/system. In the context of our discussion of functionalism, one can plausibly argue for the claim that functional properties of mental states are not independent of the physical/spiritual properties of their realizers/hosts; they actually emerge directly out of the physical/spiritual properties of their hosts. So, there is an intimate connection between the role X and the (physical/spiritual) occupant of that role X (Armstrong, 1968; Lewis, 1966).

This way of reading functionalism, I believe, may prove to be much more fruitful and rewarding. Let us turn back to the problem of intentionality for functionalism and attempt to develop a possible answer by adopting an Aristotelian framework at the background this time. One of the effective ways to describe intentionality is that intentional systems differ from non-intentional systems in that states of the former have the capacity to represent things in the world as being in a certain way, whereas states of the latter do not have any capacity to represent anything at all. At this point, the central issue of intentionality appears to be based on the concept of representation. From a broad perspective, the question of artificial and natural intelligence is really a question about the nature of representation; i.e. what is it for something to (mentally) represent something else? So, if we have a good theory of representation, we will be able to answer the question of (1) the emergence of intentionality and (2) whether it is possible for artificial systems to be capable of intentional states.

Elsewhere (in my ongoing Ph.D. thesis) I developed a theory of representation and I attempted, in particular, to answer the question of what it is to have a mental representation. From my point of view, mental representation is a special kind of structure (i.e. a map-like structure) that originates from the physical/structural resources of representing vehicles and it emerges as a result of evolutionary history of the being/system's representational mechanism. If mental representations are the product of representing vehicles and if the contents of mental representations are limited and sensitive to what physical/structural properties of representing vehicles are capable of representing plus the pressing environmental conditions determining the evolution of the function of those representing vehicles, then, it follows, two systems/beings which are unlike with respect to representational vehicles and evolutionary history cannot entertain the same representational content. For instance, a human being and a robot will never have the same mental representation of a dog, or the event of a

snake crawling on the sand. For, they will have different modes of presentations under which any object or event is conceived. The object "dog" will mean different things for them.

In conclusion, I do not claim that it is logically or technically impossible to create or build artifacts that have intelligence and psychology on their own. I do not believe, however, that there is a universal psychology that is up and running across radically different kinds of beings/systems. I attempted to show, in a number of points, where the idea of a universal psychology is very problematic. In the final analysis, it does seem to me that a human being and a robot may look at the same dog and can produce the same behavioral response, but it is highly unlikely that they will have the same representation of that dog. They do not have the same representing vehicles, nor do they have the same evolutionary history underlying the teleological function of those vehicles. Thus, neither representational content nor a universal psychology will be shared.

Acknowledgements

I owe special thanks to Prof. Stephen Voss for his invaluable criticisms and suggestions.

References

Armstrong, D. (1968). *A materialistic theory of the mind*. London: RKP.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in The Philosophy of Science*, *9,* 261–325

Block, N. (1980). Introduction: What is functionalism? In N. Block (Ed.), *Readings in philosophy of psychology*, Vol. 1, (pp. 171-184). Cambridge, Mass.: Harvard University Press.

Block, N. (1993). The computer model of the mind. In A. I. Goldman (Ed.), *Readings in philosophy and cognitive science* (pp. 819-831). Cambridge, Mass.: Harvard University Press.

Cummins, R. (1986). Inexplicit information. In M. Brand, & R. Harnish (Eds.), *The representation of knowledge and belief* (pp. 116-126). Tucson, AZ: Arizona University Press.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, Mass.: MIT Press.

Eliasmith, C. (2002). The myth of the Turing Machine: The failings of functionalism and related theses. *Journal of Experimental and Theoretical Artificial Intelligence*, *14*(1), 1-8.

Heil, J. (1998). *Philosophy of mind: A contemporary introduction*. London: Routledge.

Kim, J. (1996). *Philosophy of mind*. Boulder, Col.: Westview Press.

Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, *63*, 17–25.

Lycan, W. G. (1990). The continuity of levels of nature. In W. G. Lycan (Ed.), *Mind and cognition: A reader* (pp. 77-96). Cambridge: Basil Blackwell.

Moor, J. H. (1978). Three myths of computer science. *The British Journal for the Philosophy of Science*, *29*(3), 213-222.

Putnam, H. (1975). The meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131-193.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*, 417-57.

Shagrir, O. (2005). The rise and fall of computational functionalism. In Yemima Ben-Menahem (Ed.), *Hillary Putnam* (pp. 220-250). Cambridge: Cambridge University Press.

Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in Operations Research. *Operations Research*, *6*(1), 1-10.

Skokowski, P. (1994). Can computers carry content "inexplicitly"? *Minds and Machines*, *4*(3), 333-44.

Watson, R. A. (1995). *Representational ideas: From Plato to Patricia Churchland*. Dordrecht: Kluwer Academic Publishers.