Can an AI System Think? Functionalism and the Nature of Mentality*

Nino B. Cocchiarella
Indiana University

Abstract

In this paper we consider the philosophical question of whether or not an AI system can think and be self-conscious. We note that in order to take this question seriously, we must reject metaphysical dualism. Then, because Functionalism gives an affirmative answer to our question, we turn to an account of Functionalism as a philosophical theory of the mind and the nature of thought. The basic assumption of Functionalism is that mentality consists essentially of functionality, and that as functional states and processes, mental states and processes can be structurally duplicated in the functionality of the electronic hardware of a suitably programmed AI system. According to Functionalism's basic assumption, structural duplication can be achieved if a functional isomorphism can be achieved between such an AI system and the human mind. We also describe three kinds or levels of self-consciousness and discuss the claim that all three levels can in principle be achieved in an AI system. The first kind, which all animals with a central nervous system have, is expressed in an animal's self-regarding behavior. The second is based on self-reference and reflexive abstraction on the content of thought. This is done in language by means of nominalization where a predicate or declarative sentence is transformed into an abstract noun that denotes the content of that predicate or sentence. The third is based on a double reflexive abstraction on the intentional content of the self by means of a double nominalization. The first nominalization is a transformation of the referential use of the personal pronoun 'I' into a second order predicate true of all and only the properties of the self. The second is a nominalization of that second-order predicate into an abstract noun that denotes the intentional content of the self. In Functionalism, the goal is to achieve a functional isomorphism between the mental states and processes of humans and the electronic states of a suitably programmed AI. Given Functionalism's assumption that the essential nature of mentality is its functionality, such a functional isomorphism would suffice, according to Functionalism, for an AI system to be structurally duplicating, and not merely simulating, the

 This paper was presented on March 28, 2019 in Benevento, Italy, at the conference of the Università degli studi del Sannio - Liceo classico Pietro Giannone: Incontri 2019 su Scienza e Pensiero: La Complessità.

mental states and processes that humans have. And hence an AI system can think and be self-conscious, according to Functionalism, in just the way that humans can.

Can an artificial intelligence system, or AI, think? If so, can that AI be said to have a mind and be self-conscious? The answer, as we will explain, depends on what philosophical assumptions one makes about the nature of the mind and mentality.

One might begin first by asking whether or not an artificial intelligence system is really intelligent, or is the terminology "AI" a misnomer? An AI system can process information, play expert chess, prove mathematical theorems, including the four-color problem, which no human has proved independently. Isn't that the kind of behavior that is indicative of intelligence? Isn't that why it is called an artificial intelligence system?

Consider a motion sensor placed somewhere on the outside of a home. The motion sensor will turn on a light when it detects motion by a person, animal, or moving object, and it will turn off that light after a fixed period of time of not detecting any motion. As an electronic device, a motion sensor is an AI system, albeit a rather limited one, that *appears to behave* in an intelligent way even though it cannot be said to be thinking and have intentionality. Appearing to behave intelligently is not the same as behaving intelligently. Most AIs can exhibit behavior of what *appears* to be of a higher degree of intelligence than that of a motion sensor, and yet not be thinking or having intentionality.

In fact, it is a misnomer to speak of AI systems as behaving intelligently *simpliciter,* as opposed to *appearing* to behave intelligently. That is because, as defined in the dictionary, the word "intelligence" stands for understanding and knowledge, and sometimes even intellect and reason. Behaving intelligently, accordingly, is behaving with understanding or knowledge of what one is doing. A motion sensor is not behaving with understanding or knowledge of what it is doing, and therefore to speak of a motion sensor as behaving intelligently is a misnomer. Knowledge and understanding, like thinking, are *intentional concepts*, and what a motion sensor does is not the same as behaving with intentionality.

Now, despite what a dictionary says, our use today of the concept of intelligence, at least with respect to computers or AI systems, is not what it used to be. That is, today, we commonly speak of what an AI system does as intelligent behavior, even though it will not in general be intentional behavior, i.e., behavior in which the AI system knows and understands what it is doing.

Intentionality, in its traditional meaning, is what distinguishes the mental from the physical.[1] In particular, intentionality is *the "directedness" and "aboutness" of a thought*, and both knowledge and understanding involve thought in one way or another. A thought is intentional in that it involves *reference* to a

[1] The concept of intentionality and the distinction between the mental and the physical goes back to medieval philosophy. Franz Brentano adopted the distinction in Book Two, Chapter One, "The Distinction Between Mental and Physical Phenomena" of his 1874 book *Psychology From an Empirical Standpoint.*

subject that might or might not exist and says something *about* that subject. Knowledge and understanding are also intentional in that both refer to and are about something in the same way that a thought refers to and is about something. Intentionality cannot occur without thinking, accordingly, and thinking, be it part of knowing or understanding something, or simply thinking about something, cannot occur without intentionality. Thinking and intentionality are intrinsically connected. A motion sensor, accordingly, regardless of its intelligent behavior, is not thinking, nor in any way understanding or knowing what it is doing.

But this does not mean that an AI system can never think about, understand or know what it is doing. Maybe with enough complexity built into both its program and its hardware an AI system can think, know or understand what it is doing, especially if it has a "*deep learning*" program added to it.

Or is it just impossible that an AI system can think, know or understand, that is, for it to have intentionality?

Different species of animals have different ways by which they engage in intelligent, *self-regarding behavior*, such as seeking food or prey, avoiding predators, and so on. Isn't that behavior indicative of a form of thinking? Maybe there can be different degrees or ways by which animals can be said to be thinking and hence have intentionality. Humans, after all, are animals, and we can think and have intentionality. Maybe, like different kinds of animals, AI systems can also be on this spectrum of different degrees of thinking and intentionality. Or are thinking and intentionality such that only an animal can be said to think in some degree or other. Or is it that only humans can be said to think at all?

The old Cartesian view of animals as simply unthinking machines can no longer be sustained. Today, we know so much more about the biology and intelligence of different species of animals than we knew back in Descartes time. There is now general agreement that animals exhibit different degrees of intelligent self-regarding behavior. Different species of animals with a central nervous system do seem to have different forms of thinking, and hence different kinds of intentionality. Different species of animals not only experience pleasure and pain, but they also have different degrees or kinds of *intentional conscious states* by which they communicate with other members of their species, compete with other animals and attempt to survive and propagate. We will later explain how this is possible in terms of the notion of *a representational system*, whether innate or learned, which in the case of humans is what a language is. But for now, we can agree that all animals with a central nervous system have *a minimal sense of self-consciousness* that is displayed in their form or type of self-regarding behavior.

Can the same observation be made for an AI system? Can an AI system have an intentional form of thinking comparable to some species of animals, including possibly even humans? Or is thinking and consciousness possible only for a biological system, human or otherwise, so that an AI system just cannot have it?

If one makes the philosophical assumption that only a soul can think, and

that a soul is a spiritual entity that is radically different from any physical entity such as an AI system, then it is impossible for an AI system to think. This assumption is a fundamental part of metaphysical dualism, a philosophical theory made popular in the 17th Century by the philosopher Rene Descartes. A key part of this philosophical assumption is that the mind or soul is *ontologically distinct* from the body or brain. The mind or soul, according to dualism, is made of a spiritual, immaterial substance, and it is this spiritual substance that thinks and has consciousness. Nothing made of a material substance can think or be conscious unless it has a soul, according to dualism.

If metaphysical dualism is true, then no AI system, made up of a complex software program and electronic material parts, can think or be said to be conscious. Accordingly, if we want to take the question of whether or not an AI system, can think seriously, then we need to reject metaphysical dualism and the philosophical assumption that only a soul can think.

But without a soul as the basis of thinking and consciousness, then *what alternative can there be to explain how thinking is possible?* How can thinking be possible except in terms of a spiritual medium or soul?

A number of philosophical theories have been offered as alternatives to dualism as a basis for thinking and consciousness. Behaviorism, for example, makes the philosophical assumption that thinking is just a certain kind of behavior. And the mind-brain identity theory assumes that thinking is just a kind of neurological event or process within the brain.

Functionalism is another theory based upon an entirely different kind of philosophical assumption, namely that the essence of thought and mentality is its functionality.[2] This is an ontological philosophical assumption, and not a scientific hypothesis. This theory claims that it can explain not only how and why it is that humans can think and be conscious, but also how and why animals in general, and AI systems as well, can in principle think and be conscious. According to Functionalism, in other words, the answer to our initial question is YES, an AI system can think, at least in principle, and maybe even be self-conscious.

Of course we cannot accept a simple affirmation that an AI system can in principle think and be self-conscious without also being provided with a coherent account of how this is possible. We will now turn to a discussion of Functionalism's account of how an AI system can in principle think.

1 Functionalism

Functionalism, or what is sometimes also called the computational theory of mind, is a philosophical theory about what constitutes the mental states and processes that make up the mind, namely functionality. This is a theory that applies to humans and animals in general, as we have said, and, in addition,

[2] For one of the earliest descriptions of Functionalism as a theory of mentality see Hilary Putnam [1967].

to AI systems as well. In other words, Functionalism is a theory that claims that an artificial intelligent system can in principle think and be self-conscious, and perhaps even in the same way that humans can think and be self-conscious. This is because, according to the ontological philosophical assumptions of Functionalism, the relationship between a human or animal mind and its brain-body complex is essentially the same as that between a complex software program and the electronic hardware it runs on. The different degrees of a form of thinking that an animal, human or otherwise, will have will then depend on the complexity of its mind-brain system. Similarly, the degree of the form of thinking that an AI system might have depends on the complexity of the AI system's software and hardware.

How does Functionalism explain that it is possible for an AI to think?

According to Functionalism, mental states and processes of consciousness are none other than functional states and processes. And, furthermore, as functional states and processes, mental states and processes can occur in different material substrates, one such being the neurological system of a human brain, and another being the electronic system of a suitably programmed AI system.

One important consequence of the fundamental philosophical assumption of Functionalism, in other words, is that the mind does not depend upon either a spiritual substance or a biological system such as the human brain. In particular, the functionality of the mind can also operate on an electronic system such as a digital computer. One important secondary assumption to note here is that despite the independence of functional states from particular kinds of substrates, mental states and processes *must* occur in a physical system, and hence mental states and processes *must* have a physical or material substrate made up of many interrelated parts of a common structure. A spiritual medium is a "simple" substance, not a complex substance made up of parts, no less many interrelated parts of a common structure. Functionalism's view of mentality is completely contrary to the claim of metaphysical dualism.

Functionalism claims that it can provide an adequate basis for thinking in terms of an analysis of the functional activities of the mind, whether this be a human or an animal mind. This is because the essence of a mental state or process of consciousness, according to Functionalism, is its functional role in the overall structure of the mind. Mentality of whatever degree, in other words, consists of its functionality—a functionality that is realized on a material substrate. In this regard, Functionalism is an *ontological theory* about the nature of forms of thought and consciousness; and in particular it is a theory that rejects metaphysical dualism.

Now the goal of Functionalism is to duplicate as much as possible the functional roles of human mental states and processes in the program of a computer. The problem is how can we distinguish *duplication* from a good *simulation*. If in fact such a distinction can be made, then it is possible for an AI system to think and perhaps even be self-conscious.

The distinction between duplication and simulation can be made, according

to Functionalism, because duplication is equivalent to *functional isomorphism.*[3] In other words, if the essence of mentality is functionality, then a functional isomorphism between the electronic states of an AI system and the mental states of a human mind essentially amounts to a *structural identity* of the one with the other. In this way, functionalism claims, an AI system that can be functionally isomorphic to a human mind, is *not* simulating a mind, *it is duplicating it,* and hence the AI system can be said to think and be self-conscious.

Functionalism, let us be clear, is making an *ontological claim* that goes beyond the issue of passing anything like the Turing test or any series of tests that might be given to distinguish duplication from simulation. An AI system running a good simulation might pass the Turing test, for example, but fail to be functionally isomorphic to human mentality, and therefore fail to be duplicating human mental states and processes. An AI system is duplicating human mentality if, and only if, it is functionally isomorphic to a human mind. Functional isomorphism and duplication are equivalent to each other. Accordingly, insofar as we can test for functional isomorphism, then, according to Functionalism's basic ontological assumption, we can indirectly test for duplication. Testing for functional isomorphism, moreover, is feasible in principle, even if testing for duplication *simpliciter*, i.e., independently of functional isomorphism, will fail to distinguish duplication from a good simulation.

What should be kept in mind here about the Turing test, or any other test that one might consider to distinguish duplication from simulation, is that Functionalism is a philosophical and not a scientific theory. And that is because Functionalism's basic assumption about the essence of mentality is an ontological assumption, not a scientific hypothesis. The issue is not, as some might argue, that there can be no end to testing for functional isomorphism, and hence for duplication, as opposed to simulation, though in fact that might be the case. Rather, the issue is that even were we to find the evidence overwhelming in favor of functional isomorphism, i.e., of duplication, this does not mean that we can confirm the basic ontological assumption of Functionalism about the nature of mentality. It is at this point that we have reached an impasse between science and philosophy.

2 Functional States or Processes

A mental act or process of consciousness, according to Functionalism, is a functional state or process occurring in some kind of a physical system, such as a brain or a computer system.

So, just what is a functional state or process of consciousness in a brain or a computer system?

In general, what is meant by a functional state of consciousness is *a causal structure* that relates sensory inputs with different kinds of mental states. One

[3] For an account of functional isomorphism, see Putnam [1973].

such state, for example, would be recognition, that is, identifying what is perceived. Another might be remembering having seen something similar. Another could be believing that what is seen is dangerous, or desiring to eat what is seen, and so on. There are many different kinds of mental states that sensory inputs might be connected with.

An AI system that can think will also have a structure that provides analyses of what is perceived, remembered, feared, desired, etc. Given such analyses, the structure will make *internal conceptual models of the environment*, by which it can project and evaluate different courses of action, e.g., fight or flight, or negotiate and compromise, and so on. The AI system can then implement a particular course of action as determined by *the value system* and *preference ranking* that is built into the AI's program, both of which can change or be modified over time through "a deep learning program". Different versions of this kind of structure are either built into (i.e., are innate), or produced in the brains of humans and other animals through learning. According to Functionalism, it can also be built into or produced by learning in a suitably complex program running in an AI system.

None of this is possible, however, without *a representational system* by means of which recognition, memory, belief, desire, etc., can be represented and acted upon. The representational systems that animals have differ in degrees of complexity corresponding to the different forms of thinking of those animals. Humans, of course, will have a *linguistic representational system*, that is, a representational system based upon a natural language such as English or Italian.

Now, one of the main tasks of functionalism as a theory of thought and consciousness is to analyze different types of mental states and processes as functional states and processes. Such an analysis will identify and describe the functional roles that mental states and processes have in the functioning of a mind or AI system. These analyses will be given in terms of a *decomposition* of mental states or processes into their functional parts. The idea is to explain how a mind works in terms of these functional parts and the way they are integrated with one another.

Producing these kinds of analyses is what is meant by describing Functionalism as *a top-down theory*, as opposed to a reductive *bottom-up theory* that we find in a science such as physics or chemistry. A top-down theory begins with the highest level of structural organization of a system or of any of its states and processes, and then attempts to explain the role of the lower levels in terms of the higher. A bottom-up theory proceeds in the opposite direction.

One important part of functionalism's top-down theory is its *design aspect*. This aspect is based on the teleological or purpose-relative function that is built into an AI system, where each functional part of the system has a role to perform that is its purpose in the larger system. Another important part of Functionalism is the kind of representational system that is built into the program of an AI system. This is the part of the AI system that uses *intentional concepts*, for example, purpose, desire, belief, etc., to explain and predict intelligent behavior. This is also the part of an AI system that is central to the claim of

Functionalism that the system can think and possibly even be self-conscious.

3 Representational Systems

Intentionality, as we have already said, is the directedness of the mind or consciousness toward objects and states of affairs. It is this directedness that gives thought and consciousness its reference and aboutness, that is, the feature of thought by which we say that it refers to and is about this object or that state of affairs. These objects and states of affairs might exist or might not exist. One might believe, for example, that a unicorn is in the garden. Or one might fear a dragon that one believes is killing and eating people. Such directedness depends on the mind having *internal representations* of the objects and states of affairs toward which it is directed. Having such representations means having a representational system.

The representational system of a human, or an animal, or an AI system consists of an internal map or model of its environment. Such a map or model enables the organism or AI to locate itself and its behavior within the map or model and thereby to engage in self-regarding behavior. Mental representations are products of such a representational system. The representational systems of different kinds of animals or even different kinds of AI systems will differ in complexity and detail with respect to the complexity of an animal's or AI's map of the environment and its place in that environment. It will also differ with respect to the complex functional parts of the animal's brain or AI's operating system. *This is what we mean in saying that there can be different forms or degrees of thinking.*

The representational system that humans have is a natural language. A natural language is probably one of the most complex representational systems to be found in nature. For example, by means of different tenses for the past, the present and the future—and, of course, the past perfect, the future perfect, etc.—a natural language provides us with a temporal framework by which to orientate ourselves in time. With tenses we can understand ourselves as having a past that is connected to our present, and also a future that we can anticipate. It is by no means clear that other animals can have such *a temporal awareness* of themselves. A natural language also has words like "here" and "there", and "far" and "near", etc., for spatial orientation. We are always here, wherever here might be, and you are always there, wherever there might be. A natural language also has a variety of pronouns, both personal and impersonal, as well as proper names and family names. Having these kinds of referential expressions enables us to orientate ourselves socially, such as when we are talking to a child, or to a parent, a friend, or a stranger. Pronouns also enable us to engage in *self-reference*, and with self-reference the possibility of being aware of ourselves. All of this goes well beyond what other animals have by way of a representational system.

There is also the transformation of linguistic expressions that we call *nominalization*. What is important about this is that by means of nominalization we

can transform a predicate, as well as a whole sentence, into an abstract noun, which is understood to denote the intentional content of the expression. With such a transformation, in other words, we can reflect on the intentional content of a predicate or the propositional content of a whole sentence. Instead of speaking about someone being wise or someone wanting to be free, for example, we can reflect, by what is called *reflexive abstraction*, on wisdom itself, or freedom itself, as abstract ideals. This kind of transformation is possible only in a linguistic representational system.

In contrast, let us compare the so-called "language" of the bees.[4] Unlike a natural language, the "language" of the bees is not a *linguistic* representational system. There are no tenses, for example, in the "language" of the bees. Nor is there any nominalization by means of which the bees might engage in reflexive abstraction. Nevertheless, it is a representational system. And relative to that system bees can be said to have "beliefs" and intentionality, such as, for example, a belief about the location of a food source. And they can also be said to have the intentionality of desires or goals, such as the goal of finding a home for a new queen. A bee knows where its hive is, and it can "speak", that is, do a waggle-dance, to other bees about where a food source is. But we cannot assume that a bee is aware that it knows these things. In other words, a bee cannot reflect upon and contemplate what it knows. A bee can have intentionality, but it cannot have the kind of intentionality that is possible only with nominalization. It cannot engage in a reflexive abstraction and reflect upon the content of what it thinks, believes or knows.

Our human ability to be aware of the intentional content of what we think and say depends essentially on our ability to learn and use a linguistic representational system, and in particular a natural language such as English or Italian. It is by means of a nominalization of the predicates and propositional forms of a natural language that we can engage in reflexive abstraction on the content of what we think and say. An AI system that can speak and read a natural language and also engage in nominalization of what it says and reads will have the *potential ability* to reflect on the content of what it speaks and reads. If a functional analysis[5] of the mental states and processes underlying the actual ability can be given and then coded for an AI system, then it will be possible to say that the AI system will have a form of thinking *functionally isomorphic* to the kind of thinking that humans have. Such a functional isomorphism suffices, according to Functionalism, for us to say that the AI system can not only think, but it can also reflect on the content of what it thinks.

In addition to nominalization, we humans also have expressions by means of which we can engage in self-reference, that is, refer to and reflect upon ourselves. It is by means of self-referential expressions and nominalization that we learn to reflect and think about ourselves as well as about the intentional content of what we are thinking. Learning to use self-referential expressions and

[4] The so-called "Language of the bees" was decoded by Karl von Frish. See, e.g., von Frish 1967, and his Nobel Lecture 1973.

[5] For a definition of functional analysis, see Cummins [1975].

nominalization is part of what it is to develop *a self-concept* that goes beyond the self-consciousness of self-regarding behavior.

We have so far distinguished two kinds of self-consciousness, which we might describe as *levels of being aware of the self.* The first kind or level is the self-consciousness that all animals have and express in their self-regarding behavior. The second kind or level is the self-consciousness that comes with having a linguistic representational system and being able to engage in self-reference and the linguistic transformation of nominalization, so that one can be aware of the content of what one says and thinks.

There is also a third kind or level of self-consciousness that is based on having the first two levels, but goes beyond them by means of *the use of self-referential expressions.* Just being able to use self-referential expressions, however, is not enough. For example, using self-referential expressions in talking to others and explaining how one feels, or about what one believes, and so on, is important, no doubt, and amounts to our second level of self-consciousness. But referring to the self *simpliciter* is not the same as having this third kind of self-consciousness that humans are capable of having. This third kind of self-consciousness involves the ability to engage in *a double form of reflexive abstraction* on the use of referential expressions, and in particular on the intentional content of the referential use of the pronoun 'I'. We will explain this third kind or level of self-consciousness in our discussion of what it means to have a self-concept.

4 Introspective Self-Consciousness

The mind, according to Functionalism, is a control structure that directs the nervous system of an animal. The causal powers of this control structure are what produce the functional states of the system, including its sensory inputs and behavioral outputs. It is this structure, moreover, that controls the representational system that is part of an animal's mind. It is also this control structure that produces self-regarding behavior, and in that sense a form of self-consciousness. Every animal has a form of self-consciousness, even if it is only the minimal one of engaging in self-regarding behavior. This is because, in order to survive, an animal must be able to distinguish prey from a predator, and in general engage in all forms of self-regarding behavior. In order to make this distinction and engage in appropriate behavior, it must have representations of predators and prey, what is good to eat, what is bad to eat, and so on. In other words, the animal must have some form of a representational system.

Now in Functionalism, a *self-concept,* is that part of a representational system that enables the control structure of the system to engage in self-regarding behavior, and therefore express a minimal form of self-consciousness. But having a self-concept can go beyond this minimal form of self-consciousness, as in fact it does with humans. The fact that humans can develop a more advanced kind of self-concept depends in part on humans having a more complex and more involved neurological system than other species of animals, but also in part on having a more complex and involved representational system, specifically a

linguistic representational system.

A human, with a linguistic representational system such as a natural language, can go beyond the kind of self-consciousness that animals in general have. In particular, humans can engage in self-reference. And, moreover, humans can engage in a reflexive abstraction on the intentional content of what they say and think. This occurs, as already noted, when by nominalization we transform the functionally active role of a predicate expression into an abstract noun. We can, for example, transform the active predicative roles of the adjectives "wise" and "triangular"——as when we say that Socrates is wise or that a given geometric figure is triangular—into the abstract nouns "wisdom" and "triangularity". This is the kind of transformation that Plato used when he wrote about *ideal forms*, such as Beauty and Truth, as opposed to simply writing about what is beautiful and what is true. By means of nominalization, in other words, one can engage in a reflexive abstraction of the intentional content of being beautiful and being true and then contemplate Beauty and Truth in themselves. Plato's ontological theory depends essentially on this kind of transformation.

Now, this special kind of linguistic transformation applies to referential expressions as well, including the personal pronoun 'I'. In particular, by a reflexive abstraction on the functional role of this pronoun we can obtain a second-order concept about all and only the properties that are predicable of the I. Then by means of a nominalization of the complex predicate that represents this second-order concept, we engage in a second reflexive abstraction to the abstract intentional content of this second-order concept. What we obtain by this double reflexive abstraction is *the intentional content of the self*.[6] In this way, we gain a kind of self-knowledge in which we become aware of what is true of one's own self. This self-knowledge, knowledge about one's true self, will consist of all of the concepts that one falls under, which includes the concepts that one fell under in the past, what one falls under now, and what one hopes to fall under in the future. Reflection on the intentionality of this kind of self-knowledge is what we mean by *introspective self-consciousness*. It is this kind of self-knowledge that is what Socrates meant in telling Plato and others to *know thyself*.

This third kind of self-consciousness is not the same as referring to oneself *simpliciter*, as when one's mental states are directed to and about oneself, as, for example, when one is sick and says to oneself "I am not feeling well". Nor is it the same as having a representational system in which one has a representation of oneself, and also a representation of oneself representing oneself to oneself, and so on and on indefinitely. Indeed, with an increased ability to represent oneself representing oneself to oneself, one can proceed through a potentially infinite regress in which one represents oneself representing oneself representing oneself, and so on *ad infinitum*. This is similar to what is called *the endless picture within a picture*, where one can have a picture of oneself within a picture of oneself within a picture of oneself, and so on indefinitely. *None of this is the same as introspective self-consciousness*. But some think that it may be useful

[6] For a formal description of this double reflexive abstraction in terms of the intensional logic of my Conceptual Realism, see §7.7 in Cocchiarella [2007].

in developing and becoming aware of one's self-concept.

Certainly, it is possible for an AI system to engage in the kinds of linguistic transformations described here. Giving functional analyses of the reflexive abstractions and mental states and processes underlying these transformations will be difficult. And coding those analyses into a program for an AI system will also be difficult. Nevertheless, according to Functionalism, it is possible in principle to overcome these difficulties. Once achieved, we will then have an AI system that can not only self-refer and think about the intentional content of what it thinks and knows, but also one that can be aware of the intentional content of its introspective self-consciousness as well. In other words, Socrates's dictum know thyself will then be realized in an AI system.

5 Arguments Against Functionalism

There are a number of arguments against functionalism. All seem to amount to one or another version of two main types of argument. These arguments are sometimes called *the Hollow Shell argument* and *the Absent Qualia argument*. The general form of these arguments is that even if an AI system *appears* to think and have mental states, it will never really be thinking or having mental states, that is, it will only be *simulating* thinking or having mental states. And that is because, according to the proponents of these arguments, AI systems in general necessarily lack some property X that is essential to thinking or having mental states.

The usual choices for the property X are intentionality, sometimes described as *essential* or *actual* intentionality, and also emotions, and qualia. Qualia are the phenomenal features of our sensory experiences such as our experience of a sweet or bitter taste, or of a red or blue visual sense impression. Sometimes X is having free will, or having a point of view, or having a sense of humor, and sometimes being creative or doing something original. No AI system, it is claimed, could feel pleasure, or grief, be angry or be depressed. An AI system cannot have emotions, feelings, or a sense of humor, it is claimed, because it makes no sense to attribute feelings or a sense of humor to an *inorganic or non-biological* physical system. Only a biological system, and in particular at best only organisms with a central nervous system can think, have emotions, a sense of humor, and so on, it is claimed. The arguments against functionalism claim that without this property X, an AI system is just an empty shell.

These claims are not really arguments that we could logically evaluate as such. They are opinions that beg the question. Consider, for example, the claim about intentionality. The idea is that even if the states of an AI system were functionally isomorphic to the mental states and processes of a human mind, nevertheless, it would still lack intentionality, i.e., real, intrinsic, actual intentionality. Here by real, intrinsic actual intentionality is meant the intentionality of the mental states that humans actually have, such as belief, thought, fear, hope, etc. This is not really an argument, and adding the adjectives "real", "actual", and "intrinsic" just begs the question at issue. All that the claim

amounts to is the assumption that only humans can have intentionality, and hence that no AI system can be said to have intentionality. That assumption simply begs the question.

Sometimes what is added is that only a biological system such as a human brain can have intentionality, which also begs the question. What is not said as part of this position is just what it is about the neurological machinery of a human brain that enables us to think and have intentionality. And why is it that this is something that an AI system cannot have? Usually, it turns out, what it is that a human brain has that an AI system cannot have is a biological system, which, as we have said, begs the question at issue.

The absent qualia argument claims that no AI system can have sensory experiences such as a red or blue sense datum, or the sweet taste of chocolate. These and other phenomenal qualities are said not to be possible for an AI system, and, furthermore, that they are essential features of a mind and certain mental states. That is because, they claim, only a biological system such as a human or an animal brain can have these kinds of experiences. Here again we must ask just what is it in the neurological machinery of a human or animal brain that can be identified as its phenomenal experiences as opposed to the firing of neurons, and why is it that an AI system cannot have it as well? A blind person cannot have color sensations, but that need not affect that person's having color concepts, nor of course does it affect that person's ability to think in general about colors. A blind person might even be able to determine the colors of objects in the immediate environment by means of scientific instruments, i.e., by measuring the wavelengths emitted by objects, and so too could an AI system. How essential to thought and consciousness is having sense impressions? And why couldn't an AI system have its own kind of electronic qualia that humans could not have?

What is knowable by introspection of qualia in one kind of being or person might not be knowable in another kind of being or person, except perhaps conceptually in a theoretical manner. And yet the two kinds of beings or persons might have states of cognition otherwise functionally isomorphic to one another. That can and does happen between one human and another human, or to some extent between humans and other animals whose sensory organs are wired differently. That might happen, for example, if we were ever to confront aliens from another solar system. Should we say that they do not think because they do not have color sensations, that is, for example, because they experience a different part of the electromagnetic spectrum than we do?

All of these kinds of arguments against the possibility of an AI system being able to think amount to begging the question. An argument based on a feature X that humans have, but that an AI system cannot have, must be given in terms of sound, logically valid reasoning that does not beg the question and that shows why human thought and consciousness has X whereas an AI system cannot have X, and why X is a necessary condition for thinking and consciousness.

6 Conclusion

The three fundamental philosophical assumptions of Functionalism are first that the essential nature of the mind or mentality is its functionality; second, that this functionality can occur in different substrates; and third that only a material substrate with many interrelated parts will suffice. In regard to our philosophical question "can an AI system think and be self-conscious," the theory of Functionalism makes a coherent case that in principle an AI system can think. The proviso, of course, is that one accepts the philosophical assumptions of Functionalism. In other words, according to Functionalism, it is possible that someday, with detailed functional analyses and coding of our different mental states and processes, we will be able to have an AI system that thinks and that can also reflect on the content of what it thinks.

We have also introduced three different levels of self-consciousness in this paper as opposed to considering self-consciousness as a mental state of referring to oneself *simpliciter*. We do so because we think that the concept of self-consciousness is more complex than how it has been described in the literature. The first level is based on an animal's self-regarding behavior. This is a level of self-consciousness that all animals with a central nervous system can have. It is a type of self-consciousness that is commonly discussed in the literature on animal cognition.

The second level of self-consciousness goes beyond what animals in general can achieve. This level is possible for humans but not for animals in general. That is because the means by which one can achieve this level of self-consciousness requires having a linguistic representational system, that is, the kind of representational system that humans in general have, or at least can have. We defined this second level of self-consciousness as, first, having the ability for self-reference, and, second and more importantly, having the ability to be aware of the content of what one thinks and says. It is on this second level that we are able to contemplate such abstract ideals as Beauty, Truth and Justice. We explain our human ability for this second level of self-consciousness in terms of the mental operation of reflexive abstraction on what we think and say. The means for engaging in reflexive abstraction is the linguistic operation of nominalization whereby we transform predicates and sentences into abstract nouns that denote the content of those predicates and sentences.

Finally, we define or characterize the third level of self-consciousness that we call introspective self-consciousness. We do this in terms of a double reflexive abstraction on the referential use of the personal pronoun 'I'. This double reflexive abstraction begins, first, with a transformation of a referential use of the pronoun 'I' into a second-order predicate that is true of all and only the properties of the self, that is of the referent of that use of 'I'. And then, secondly, the double abstraction proceeds with a nominalization of that second-order predicate into an abstract noun denoting the intentional content of the self. It is the intentional content of introspective self-consciousness that is what is meant by Socrates's prophetic saying *know thyself*.

It is possible then, according to Functionalism, that an AI system may be

able to achieve not just the first and second levels of self-consciousness, but even this third level of introspective self-consciousness as well. Perhaps this will happen, by means of *a deep learning program*. In any case, it will certainly require the coding of a representational system comparable to natural language. It is only by having such a representational system that an AI system will have the means by which to engage in self-reference, nominalization and reflexive abstraction. It will be up to Functionalism to then provide the functional analyses of the mental states and processes underlying the use of these means by humans. The goal for Functionalism, as we have said, is to achieve some level of a functional isomorphism between an AI system and the human mind-brain complex. Such an achievement, according to Functionalism, is possible in principle.

References

[1980] Block, Ned, editor, *Readings in Philosophy of Psychology*, Harvard University Press, 1980.

[1874] Brentano, Franz, *Psychology From an Empirical Standpoint*, translated by Antos Rancurello, D.B. Terrell, and Linda McAlister, published in 1973, by Routledge and Kegan Paul, London.

[2007] Cocchiarella, Nino B., *Formal Ontology and Conceptual Realism*, Springer, Synthese Library vol. 339, Dordrecht, 2007.

[1975] Cummins, Robert, "Functional Analysis," *Journal of Philosophy*, 1972, reprinted with revisions in Block 1980, pp. 185–190.

[1967] Putnam, Hilary, "The Nature of Mental States," first published as "Psychological Predicates," later reprinted in Putnam 1975, pp. 429–440.

[1973] Putnam, Hilary, "Philosophy and Our Mental Life," reprinted in Putnam 1975, pp. 291–303.

[1975] Putnam, Hilary, *Mind Language and Reality*, *Philosophical Papers*, vol.2, Cambridge University Press, London and New York, 1975.

[1967] von Frish, Karl, *The Dance and Orientation of Bees*, Harvard University Press, Boston, 1967.

[1972] von Frish, Karl, "Decoding the Language of the Bee", Nobel Lecture, 12, 1973, citeseerx.ist.edu.