Mabaquiao, Napoleon. 2011."Computer Simulation of Human Thinking: An inquiry into its Possibility and Implications." *Philosophia: An International Journal of Philosophy* 40(1): 76-87. (Original pages in < >.)

COMPUTER SIMULATION OF HUMAN THINKING: AN INQUIRY INTO ITS POSSIBILITY AND IMPLICATIONS[1]

Napoleon M. Mabaquiao Jr.
De La Salle University,, Manila

Critical in the computationalist account of the mind is the phenomenon called computational or computer simulation of human thinking, which is used to establish the theses that human thinking is a computational process and that computing machines are thinking systems. Accordingly, if human thinking can be simulated computationally then human thinking is a computational process; and if human thinking is a computational process then its computational simulation is itself a thinking process. This paper shows that the said phenomenon—the computational simulation of human thinking—is ill-conceived, and that, as a consequence, the theses that it intends to establish are problematic. It is argued that what is simulated computationally is not human thinking as such but merely its behavioral manifestations; and that a computational simulation of these behavioral manifestations does not necessarily establish that human thinking is computational, as it is logically possible for a non-computational system to exhibit behaviors that lend themselves to a computational simulation.

INTRODUCTION

In addition to our physical activities, we go about our everyday lives engaging in mental activities such as thinking, making decisions, wishing, imagining, and experiencing pains and pleasures. As we attribute our physical activities to the workings of our bodies, we attribute our mental activities to the workings of our minds. But as we continue to know a great deal about our bodies and their physical environment, our minds remain to be a mystery. Yet our minds form an important part of what defines who we are as humans and as individual persons, for it is essentially in reference to our minds that we distinguish ourselves from the rest of nature and from one another. If we are to <76> fully understand who we are and our proper place in nature, we therefore need to understand the nature of our minds.

It used to be held that science could afford to ignore the human mind in its investigations, mainly believing that it is a subject matter better left to the philosophers and theologians. But somehow the idea that science can never be complete or comprehensive in its account of nature unless it also deals with the nature of the human mind persists. And so there have been natural scientists, mathematicians, and philosophers, who have attempted to come up with a scientific account of the workings of the human mind. Initial formulations of this account were met with doubts and skepticism, but with the entry of computer technology into the scene, things have taken a revolutionary turn. With its enormous capacity for simulating a wide variety of phenomena, the dreamt-of science of the mind is then believed to be already within reach.

The use of computers in understanding the human mind has given rise to the *computational theory of mind* or *computationalism* for short, according to which the human mind is a computational system or a kind of computer, and that the computer itself is a kind of mind. In the field of computer science, this theory has been labeled Strong Artificial Intentelligence or *Strong AI* for short. The projected science of the mind grounded in this theory has come to be known as *Cognitive Science*. The popularity of this theory extends from academic circles to the entertainment business, as shown by the proliferation of sci-fi movies whose themes often revolve around the possibility of creating machines that would eventually evolve into conscious beings and thereby acquiring the capacity for making autonomous decisions and for experiencing complex emotions like love, anguish, and the fear of death.

The computationalist argument, in the main, comes down to the following. If human thinking can be simulated computationally then human thinking itself is a computational process. This establishes the view that the human mind is a sophisticated type of computer. Now if human thinking is a computational process then the computational simulation of human thinking itself thinks.[2] And this establishes the view that the computer, with the appropriate degree of sophistication, possesses a mind. These considerations show how critical is the possibility of a computational simulation of human thinking in establishing the theses of computationalism.

In this paper, I will argue that the phenomenon called “computer simulation of human thinking” is ill-conceived, and that, as a consequence, what it hopes to accomplish—the theses of computationalism—are problematic. I will argue that it is not human thinking as such that is simulated computationally but its behavioral manifestations or outputs; and that a computational simulation of the behavioral manifestations of human thinking does not necessarily establish that human thinking is computational, as it is logically possible <77> for a non-computational system to exhibit behaviors that lend themselves to a computational simulation. My discussion is divided into three parts. The first looks into the competing viewpoints on the nature

of the mind in relation to its possible computational simulation. The second inquires into what is really being simulated computationally when we speak of a computational simulation of human thinking. And the third probes into the type of simulation that takes place in a computational simulation of human thinking.

THE ALTERNATIVE VIEWPOINTS

In his book, *Shadows of the mind* (1994, 12), Roger Penrose lists four alternative positions on the relation of human thinking to its possible computational simulation:

A. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.
B. Awareness is a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.
C. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be simulated computationally.
D. Awareness cannot be explained by physical, computational, or any other scientific terms.

Each of these alternative positions represents a particular viewpoint regarding the nature of the mind. Position A is the viewpoint of strong AI, which is the main proponent of computationalism in the field of computer science. This viewpoint, as earlier pointed, regards human thinking as a computational process. More precisely, it looks at the mind-brain relation as a species of the software-hardware relation in that the mind is to software as the brain is to hardware. Consequently, computers, with the appropriate degree of sophistication, are regarded by this viewpoint as thinking systems or as machines possessing minds.

Position B is the viewpoint that Penrose attributes to weak AI and, to some extent, to the view of John Searle (see 1980, 1990, 1994, 1999, and 2004) as embedded in his popular Chinese room argument. In contrast to strong AI which claims that the human mind is literally a computer, weak AI merely claims that the computer is a powerful tool for understanding the workings of the human mind. On the other hand, the said view of Searle refers to his claim that while it can be granted that human thinking is a computational process it is not all there is to human thinking; for human thinking, for Searle, is also inherently intentional, which a mere computational system, such as the computer, is not. <78>

Position C is the viewpoint that Penrose himself endorses. For Penrose, human thinking is not a computational process, which he proves by means of Gödel's

incompleteness theorem. Accordingly, since a computational system cannot prove the truth of certain propositions that are outside its system while human thinking can—as demonstrated by Gödel's incompleteness theorem, it follows that human thinking is not a computational system. Now as human thinking is not a computational system, then it cannot, for Penrose, be simulated computationally. Despite being non-computational, Penrose, however, maintains that human thinking is still a physical process and hence in principle can be accounted for by science, for he believes that human consciousness in general is a result of the quantum activities in the brain's microtubules. Quantum mechanics is here regarded by Penrose as a physical and scientific theory that is non-computational.

Position D is the viewpoint of the Cartesian dualism. The very reason why human thinking cannot be simulated on this view is that it regards the mind as something metaphysical and hence not within the grasp of science.

A critical assumption can be observed in all these positions, namely, that a computational system can be simulated computationally while a non-computational system cannot be simulated computationally. This grounds the reasoning that if a system can be simulated computationally then it is itself a computational system. Positions A and B claim that human thinking can be simulated computationally for human thinking is a computational process. Again, the difference between positions A and B is simply that position A believes that being a computational process is all there is to human thinking such that the computational simulation of human thinking itself thinks, while position B believes that there is more to human thinking than being a computational process—its inherent intentionality—such that the computational simulation of human thinking does not itself think. Positions A and B, however, both claim that human thinking is a physical process and hence can be accounted for by science.

On the other hand, positions C and D claim that human thinking cannot be simulated computationally for human thinking is not a computational process. Again the difference between positions C and D is that position C claims that human thinking is nonetheless a physical process and hence can be accounted for by science, while position D claims that human thinking is a non-physical process and hence cannot be accounted for by science. In the following discussions, we shall look more closely into this assumption.

THE OBJECTS OF SIMULATION

In the course of expounding the position that he endorses, Penrose gives three formulations of his position which point to three aspects of human thinking that he claims cannot be simulated computationally, namely: (1) thinking (understanding or consciousness) itself, (2) the <79> outward effects or the behavioral manifestations

of thinking, and (3) the physical actions of the brain that evoke thinking. Penrose fails to clarify the relations among these three aspects of human thinking, whether, for instance, he regards them to be identical or as logically implying one another. In any case, Penrose's three formulations of his position speak of three types of computational simulation involving human thinking whose relations with one another need to be clarified. These types of computational simulations are as follows:

CS1: The computational simulation of thinking
CS2: The computational simulation of brain's physical actions that evoke thinking
CS3. The computational simulation of the behavioral manifestations of thinking

The first question that we need to raise is whether these three types of computational simulations, or any two of them, are identical to one another. To begin with, CS2 and CS3 obviously cannot be identical nor can CS2 can be regarded as a sub-class of CS3. The simple reason is that brain activities are events involving the neurons while behaviors are physical movements involving the external parts of the body (see Jaegwon Kim 1998, 29).

Consequently, the possibility that CS1, CS2, and CS3 are all identical is likewise ruled out. And thus we are left with the possible identity between CS1 and CS2 on the one hand, and between CS1 and CS3 on the other. Now, to say that CS1 is the same as CS2 is to adhere to the (mind-brain) identity theory which claims that mental states are nothing but brain states; while to say that CS1 is the same as CS3 is to adhere to behaviorism (here regarded in its radical form), which claims that mental states are nothing but behaviors or, more specifically, behavioral dispositions. These two theories, the identity theory and behaviorism, however, are rejected by computationalism. Accordingly, mental states as computational states are higher-level physical states which are not reducible to either the physical states of the brain or of the body, which are the lower-level physical states. And since what is at issue are the claims of computationalism, we then have to disregard in our analysis the possible identity between CS1 and CS2 on the one hand, and between CS1 and CS3 on the other.

The next question that we need to deal with is: If not by identity then what kind of relation exists, on the one hand, between CS1 and CS2, and on the other, between CS1 and CS3? Let us begin with the consideration on whether CS1 can be established directly. If we suppose that it can be established directly then CS2 and CS3 simply become superfluous. For why would we still want or even bother to simulate computationally the physical actions of the brain that cause understanding or the behavioral manifestations of understanding when <80> what we really want to

understand is the nature of understanding itself which we can computationally simulate directly?

Granted that CS1 cannot be established directly then the next question is in what way are CS2 and CS3 relevant in establishing CS1. Roger Schank (1984, 53-54) argues that only CS3 is relevant in establishing CS1. Schank (1984, 39) explains that examining the inner processes of a system that allegedly enable such system to produce outputs to which understanding can be ascribed is either extremely difficult to do or does not really establish much. It is extremely difficult to do because if one's brain is cut open so that what is happening with its neurons while one is having some mental states can be directly observed or so that one can establish a direct correlation between a particular physical activity of the brain and a particular mental state, the brain will most likely be damaged or worse it will cease to function. But even if assuming that a direct observation of the inner processes can be done, Schank argues that it does not really say much. Why? It will be observed that a given behavior can be directly correlated to some mental state. For instance, a certain behavior can be said to be a manifestation of pain. A certain brain state, however, cannot be directly correlated to a particular mental state, for it requires the mediation of some form of behavior. For instance, if we simply know that some neurons in our brain are firing, without correlating them with some pain behavior of ours, we would not know that such brain activity is about some pain of ours.

Schank (1984, 55) further elaborates that the situation is no different from the one in which we are to determine whether extraterrestrials are capable of understanding or thinking. Since we do not know anything about their physiology then no amount of examining their "insides" (for all we know they may not even have brains that more or less resemble ours) would tell us that they are capable of understanding. The only way, in this regard, to determine whether they are capable of understanding is to examine their behaviors.

Schank's point actually just follows Alan Turing's in his imitation game. Turing's imitation game, popularly known as the Turing test, makes a certain type of human behavior to which thinking is naturally attributed as the criterion for determining whether machines can also be said to be intelligent. Such type of behavior involves providing answers to certain questions (Turing 1995, 25). The task of the machine is to mimic this type of behavior in ways that the interrogator would not be able to distinguish between the machine and the human. Accordingly, if the interrogator cannot distinguish between the human and the machine on the basis of their answers to the questions thrown to them, then the machine is said to pass the Turing test and hence can be said to be capable of thinking. Here, in order to separate the irrelevant features of the machine and human respondents or those features of theirs that will make the interrogator's judgment unfair, the interrogator <81> is separated from the human and machine respondents by a wall wherein the

interrogator's only access to the respondents is through their textual communication via a teletype machine. Thus, the interrogator has no access to the voice and other external physical features of the respondents.

The point is that the relevant consideration for the question on whether a machine thinks, in the context of the imitation game, is whether this machine can mimic the thinking behavior of the human. Consequently, whether the machine is run by a computer program or by some other mechanism, is not relevant to the question. If a machine that is run by a computer program passes the Turing test, the fact that it is run by a computer program is not what is relevant; what is relevant is whether the machine is able to exhibit outputs that mimic the thinking behaviors of the human.

VARIETIES OF COMPUTER SIMULATION

Our investigations thus far have led us to the conclusion that human thinking is never directly simulated computationally. What are directly simulated computationally are merely the behavioral manifestations or outputs of human thinking, which we can simply refer to as *thinking behaviors*. That being the case, our inquiry is thus reduced to the following questions:

1. Does the computational simulation of the thinking behaviors of humans imply that human thinking itself is computational?
2. Does the computational simulation of the thinking behaviors of humans imply that the simulating system itself thinks?

Let us again consider the Turing test. The point of Turing is merely consistency in the attribution of thinking. The reasoning of Turing is that since humans are said to be thinking when they behave in certain ways, a machine that behaves in similar ways should therefore be said to be thinking as well without prejudice to its being a machine. In this regard, the conclusion that the machine thinks is not arrived at via the fact that the machine's internal mechanism or processes simulate the internal processes of human thinking, but via the fact that the machine's external or behavioral outputs simulate human thinking behaviors. Thus we are here talking about a simulation that takes place not between the internal mechanisms of two systems or of the internal mechanism of a simulating system and the external outputs of a simulated system, but between the external outputs of two systems.

To get a grip of the implications of the consideration made above, let us examine certain types of computer simulation or various forms that a computer simulation can assume. Consider the difference between <82> a computer simulating how two persons played a chess game and a computer simulating how a human would play chess by actually playing chess with a human. In the former case, the simulation just

replicates a chess game that was already played and as such can be called a *purely representational simulation.* In the latter case, however, the simulation takes place while the computer interacts with a human or responds to some external stimuli. Here, we do not know in advance how the game will exactly be played. We can call this type of computer simulation, to distinguish it from the former one, an *interactive simulation*.[3] The relationship between these two types of simulation is that an interactive simulation is necessarily a representational simulation as well for representation is the most basic level of simulation, but a representational simulation need not be interactive as well.

Consider another difference. It is different when a computer simulation of some human physical movement can only be viewed on the computer monitor and when a computer simulation of the same is carried out by some mechanical system, say a robot. In the former we just have a graphical representation of the human physical movement that we can view on the computer monitor. In the latter, it is the physical movement of the machine run by a computer program that simulates the same human physical movement. Here, we can distinguish between two types of computer simulation: the simulation that takes place merely on the level of the computer monitor, which we can refer to simply as *graphical simulation;* and the simulation that takes place on the level of the physical movements of a machine run by a computer program, which we can refer to as *mechanical simulation.* A graphical simulation, of course, need not translate into a mechanical one. While, for instance, a graphical simulation of the weather can, in principle, be done, a mechanical simulation of it will be extremely difficult, if not altogether physically impossible, to do.

Now, in light of the previous distinction between purely representational and interactive simulation, we can thus have, on the one hand, a graphical simulation that can either be merely representational or interactive; and on the other hand, a mechanical simulation that can likewise either be merely representational or interactive. Let us take the case of a computer simulation of the human heart. Such simulation can take place only on the level of the computer monitor wherein one can view how the human heart functions. This is a graphical simulation of the human heart, which can be merely representational, wherein the program allows the user to use certain commands to easily jump from one image or piece of information about the human heart to another. But it can also be interactive in that it allows the user to interact with the program, say the user can input certain conditions in order to determine how the human heart will react. On the other hand, the simulation of the human heart can also involve a mechanical entity that simulates the physical movements of the actual <83> human heart. This then is a mechanical simulation of the human heart, which can be called a “mechanical heart”. This mechanical heart may be such that it only mimics the physical movements of the human heart but does

not actually function as a heart in the sense that it actually pumps actual human blood. In this case, such mechanical simulation is a purely representational simulation. But if it actually functions as a heart in the sense that it actually pumps actual human blood then such mechanical simulation is interactive as well.

An important thing to consider here is that in creating a program that would enable a machine to simulate human behaviors representationally or interactively, graphically or mechanically, the programmers are not really designing the program by simulating computationally such behaviors. They are designing the program, or the appropriate computations, that would enable the machine that it runs to exhibit outputs that would simulate human behaviors. In this regard, the types of computer simulations that we have considered thus far are actually just variants of a more general kind of simulation which we can refer to as *behavioral simulation*. As the term implies, a system behaviorally simulates another system when the former system behaves in ways that simulates the behaviors of the latter system.

But a behavioral simulation need not be run by a computer program. When humans, for instance, behaviorally simulates the movements of fellow humans, say a child mimicking the mannerisms of his or her parents or teachers, or the physical movements of certain animals, as when children play the role of animals in a play, what we have is a type of simulation that is not or need not be run by a computer program. Other examples include aliens or extra-terrestrials behaviorally simulating the physical movements of humans, and some physical movements in nature happen to simulate human behaviors, say a group of clouds or stars simulating human figures in motion. We cannot say that the behavioral simulations that take place here are run by computer programs. To distinguish between these two types of behavioral simulation, we can refer to the type that is run by a computer program as *computational behavioral simulation*, whereas we can refer to that type that is not run by a computer program as *non-computational behavioral simulation.*

If a behavioral simulation simulates the actual functions of what is being simulated, we can call this type of simulation a *functional simulation*. For instance, if a mechanical simulation of the human heart is able to function like a human heart in the sense that it can replace the human heart then such simulation is a functional simulation. If not, then it is merely a behavioral simulation. Note that a functional simulation is necessarily interactive but may be graphical or mechanical depending on the nature of what is being simulated. A functional simulation of the human heart is necessarily interactive and mechanical. A functional simulation of the human heart as seen only on a computer monitor does not really simulate the actual functions of the human heart. <84> On a computer monitor the simulated human heart only pumps simulated blood. But a mechanical heart can simulate the actual functions of a human heart in that it can be made to pump actual blood. In contrast, a functional simulation of human thinking is necessarily interactive but may be graphical or mechanical.

When a computer, for instance, simulates the human activity of playing chess, the simulation must be interactive but it may be graphical, as when the computer plays the game of chess with a human via the computer monitor, or mechanical, as when a robot run by a computer program plays with a human. For the human activity of answering questions, Turing claims that all that is required is that its simulation is interactive and it is irrelevant whether it is graphical or mechanical.

There are, of course, other possible types of computer simulation but the differences that we have identified will do for our purposes. These distinctions will help clarify what is really involved in certain arguments used both to challenge and defend the claims of computationalism. But here is the critical question: Is a functional simulation *necessarily* a computational simulation? Another way of putting this is: Is functional simulation necessarily run by a computer program? In so far as a behavioral simulation need not be computational, so is a functional simulation. To demonstrate this, let us consider the following thought experiment, which is actually just an extension of Schank's about the extra-terrestrial. Let us take the case of Superman, one of the popular fictional superheroes. Story has it that Superman is an alien who disguises as a human in the person of Clark Kent. Considering his superpowers, some of which are his abilities to fly, lift extremely heavy objects, see through solid objects, cut objects through his laser vision, and throw heavy objects using just his breath, and a strange cause of weakness, namely exposure to kryptonite, his whole body must surely be made of stuff very much different from what ours are made of, though from the outside it surely looks like that of a human. Just think of the fact that bullets cannot even scratch his skin though it certainly looks like ours. Likewise his "brain" or "mind" or that which somehow corresponds to what we call such must also be different in composition, structure, and functions from ours. But Superman does not only look like a human being. He is also capable of behaving as a normal human being when he plays the role of Clark Kent by intentionally committing some of the follies and manifesting some of the weaknesses humans are naturally prone to. We can thus meaningfully say that Superman as Clark Kent can easily pass the Turing test, as evidenced by his success in making most people around him believe that he is a normal human being.

Let us suppose that in the imitation game of Turing, aside from the human and machine respondents we add an alien respondent in the person of Superman playing the role of Clark Kent. With his superpowers, Superman can easily pass the Turing test along with the machine. Consequently, we should regard both machine and Superman, <85> along with the human, as thinking or intelligent entities. But can we conclude here that all these respondents, human, machine, and Superman, have the same internal mechanisms that enable them to pass the test, say that they are all, like the machine, computational entities? We cannot say it of Superman, but the same goes with the human respondent.

The whole point of the exercise is to show that a non-computational system can perfectly behave in ways that can be simulated computationally. Thus, the fact that a particular behavior can be simulated computationally does not necessarily establish that such behavior is an output of some computational system.

CONCLUSION

Let us now consider how our investigations bear directly on the questions we earlier posed; namely, does the computational simulation of the thinking behaviors of humans imply that human thinking itself is computational and does the computational simulation of the thinking behaviors of humans imply that the simulating system itself thinks? With regard to the first question, the answer is in the negative; for as we have shown it is logically possible for a non-computational system to exhibit behavioral outputs that lend themselves to computational simulations. With regard to the second one, however, the answer needs qualifications. We have shown that the attribution of thinking to a simulating system is simply based on its functional simulation of the thinking behaviors of humans without regard to the type of internal mechanism that enables the simulating system to perform the said functional simulation. In other words, the computational or non-computational nature of the simulating system is irrelevant to the attribution of thinking to the said system. Thus, if the computing machine, by means of its computational nature, is able to functionally simulate human thinking behaviors then thinking can be attributed to this machine. But its computational nature has nothing to do whatsoever with such attribution, and hence there is nothing here that will allow the inference that thinking is basically a computational process. Given this qualification, the answer then to the second question is in the affirmative.

However, if what the question intends to establish is that in simulating human thinking behaviors the simulating computational system thinks in the very same way that humans do, the answer to this question then is in the negative. And this is so in virtue of the same reason why we have given a negative answer to the first question. Machines, like humans, can be said to be thinking, but thinking as defined merely in a functional way, that is, as having the capacity to perform certain functions such as those measured in the Turing test. But in the case of humans, thinking is more than just performing certain functions. <86> Thinking, in the case of humans, is necessarily a conscious activity, which happens to elude any form of computational simulation.

NOTES

1. This paper was read at the Ariston Estrada Seminar Room of De La Salle University-Manila on 4 April 2009 in fulfillment of the requirements for the Emerita Quito Distinguished Professorial Chair in the History of Thought.
2. In this essay, we follow the practice of using the expressions "computational simulation" and "computer simulation" interchangeably. This practice is based on the view, called the *Church-Turing Thesis*, which states that what is computational is what can be implemented in a Turing machine, which is widely regarded as the abstract model of the modern-day digital computer.
3. The term "interactive" was suggested to me by Mr. Jeremiah Joven Joaquin.

REFERENCES

Kim, Jaegwon. 1998. *Philosophy of mind.* Colorado: Westview Press Inc.

Penrose, Roger. 1994. *Shadows of the mind: A search for the missing science of consciousness*. Oxford: Oxford University Press.

Schank, Roger with Peter Childers. 1984. *The cognitive computer.* Reading: Addision-Wesley Publishing Company, Inc.

Searle, John.1980. Minds, brains, and programs. Available at http://www.cogsci.soton.ac.uk/bbs/Archive/bbs.searle2.html. Accessed: 30 October 2001.

_____. 1990. Is the brain a digital computer? Available at http://cogsci.soton.ac.uk/~harnad/Papers/Py104/searle.comp.html. Accessed: 5 December 2001.

_____. 1994. *The rediscovery of the mind.* Massachusetts: The MIT Press.

_____. 1999. *Mind, language and society: Doing philosophy in the real world.* London: Weidenfeld and Nicolson.

_____. 2004. *Mind.* Oxford: Oxford University Press.

Turing, Alan. 1995. Computing machinery and intelligence. In *Computation and intelligence: Collected readings.* Edited by George Luger. Cambridge: The MIT Press.

_____. 1936. On computable numbers with an application to the entscheidungsproblem. Available at http://www.thocp.net/biographies/papers/turing_oncomputablenumbers_1936.pdf. Accessed: 31 October 2008.

Submitted: 1 May 2009; revised: 5 December 2010. <87>