Self-consciousness Cannot Be Programmed

Jinchang Wang
School of Business, Richard Stockton College of New Jersey, Galloway Township, NJ 08205, U.S.A.

Keywords: Artificial Intelligence, Self-consciousness.

Abstract: We investigate the issue about whether a computer can be self-aware or self-conscious. We derive logically that if a machine can be copied or duplicated then it cannot be self-aware. Programs of a digital computer are copiable, therefore self-consciousness cannot be programmed. Self-awareness is an insurmountable stumbling block for a digital computer to achieve full range of human consciousness. A robot cannot be self-conscious unless it is not copiable.

1 KURZWEIL'S THOUGHT LAB

We investigate the issue of machine self-consciousness and give a logical answer in this article. We start with Ray Kurzweil's thought lab.

When talking about a computer's self-identity, Kurzweil used a thought lab of copying himself with reverse engineering. *Reverse engineering* refers to replicating something by scanning its composition and structure to the levels of neural cells, molecules and atoms, and rebuilding a copy of it according to the scanned information. "If we scan – let's say myself – and record the exact state, level, and position of every neurotransmitter, synapse, neural connection, and every other relevant detail, and then reinstantiate this massive data base of information (which I estimate at thousands of trillions of bytes) into a neural computer of sufficient capacity, the person that then emerges in the machine will think that he is (and had been) me. He will say 'I grew up in Queens, New York, went to college at MIT, stayed in the Boston area, sold a few artificial intelligence companies, walked into a scanner there, and woke up in the machine here." (Kurzweil, 2002, p.42)

"Is the person emerging in the machine Ray Kurzweil?" asked Ray Kurzweil. "Objectively," Kurzweil answered, "the newly built 'Ray' is exact me in the eyes of everyone except for me." "But wait. Is this really me? For one thing, old biological me still exists. I'll still be here in my carbon-cell-based brain. Alas, I will have to sit back and watch the new Ray succeed in endeavors that I could only dream of." (Kurzweil, 2002, p.42) "If you then come to me, and say, 'Good news, Ray, we have successfully reinstantiated your mind file, so we won't be needing your old brain anymore,' I may suddenly realize the flaw in the 'identity from pattern' argument. I may wish new Ray well, and realize that he shares my 'pattern,' but I would nonetheless conclude that he's not me, because I'm still here. How could he be me? After all, I would not necessarily know that he even existed." (Kurzweil, 2002, p.43)

Kurzweil continued his thought lab in his 2005 book <Singularity is near>. "Although the copy share my pattern, it would be hard to say that the copy is me because I would - or could - still be here. … Although he would have all my memories and recall having been me, from the point in time of his creation Ray 2 would have his own unique experience, and his reality would begin to diverge from mine." "If we copy me and then destroy the original, that's the end of me, because as we concluded above the copy is not me." (Kurzweil, 2005, p.384)

Kurzweil raised a dilemma: the copy of "me" is not "me"! It is an enlightening puzzle. It leads us to think deeply into the issue of what on earth "self" is, how "myself" after copying becomes another "self", and whether an artificial robot can be spiritual. But Kurzweil did not go further and pursue to solve this dilemma.

Wang J..
Self-Consciousness Cannot Be Programmed.
DOI: 10.5220/0005285905770581
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (ICAART-2015), pages 577-581
ISBN: 978-989-758-074-1

2 IF A MACHINE IS COPIABLE THEN IT CANNOT BE SELF-CONSCIOUS

We argue in this section that a copiable artificial machine cannot be of self-awareness or self-consciousness, therefore it does not have self-identity. The arguments are logically straightforward.

2.1 Self, Self-awareness, and Their Features

According to Wikipedia, "the *self* is the subject of one's own experience of phenomena: perception, emotions, thoughts. The self is seen as requiring a reflexive perception of oneself, the individual person, meaning the self is an object of consciousness." (Wikipedia, 2014 (1))

Philosophers and psychologists view the self differently. "The philosophy of self seeks to describe essential qualities that constitute a person's uniqueness of essential being."(Wikipedia, 2014 (1)) "The psychology refers to the cognitive and affective representation of one's identity or subjective experience." (Wikipedia, 2014 (1))

"*Self-awareness* is the capacity for introspection and the reflective ability to recognize oneself as an individual separate from the environment and other individuals." "Self-awareness or self-consciousness is a form of intelligence which is an understanding of one's own existence." Similarly, "*self-identity* is an awareness of the identification with oneself as a separate individual, or the conscious recognition of the self as having a unique identity." (Wikipedia, 2014 (2))

Self and self-awareness are related. The self is a being of an entity's (or agent's) subjective phenomenon which includes one's emotions, perception, thoughts, and the self exists for the entity only if the entity is self-aware, which is a reflexive and retrospective capability to recognize the subjective phenomenon.

There are numerous definitions and discussions on what self, self-awareness, and self-identity are. We do not intend to pursue the exact definitions of them in this article. What we need here for the purpose of showing the possibility of artificial self-awareness are just some commonly accepted features of self and self-awareness.

One basic feature of self is *subjectivity*. The self refers to the first person "I". As put by Kurzweil, "When people speak of consciousness they often slip into considerations of behavioral and neurological correlates of consciousness (for example, whether or not an entity can be self-reflective). But these are third-person (objective) issues and do not represent what David Chalmers calls the 'hard question' of consciousness: how can matter (the brain) lead to something as apparently immaterial as consciousness?" (Kurzweil, 2005, p.385) "The essence of consciousness is *subjective* experience, not objective correlates of that experience." (Kurzweil, 2002, p.44) He further pointed out the un-measurableness of subjective experience, "Science is about objective measurement and logical implications therefrom, but the very nature of objectivity is that you cannot measure subjective experience – you can only measure correlates of it, such as behavior (and by behavior, I include the actions of components of an entity, such as neurons). This limitation has to do with the very mature of the concepts 'objective' and 'subjective'. Fundamentally, we cannot penetrate the subjective experience of another entity with direct objective measurement." (Kurzweil, 2002, p.45) For a particular person, there are many "himself's" and "yourself's", but there is only one "myself" which is the subjective self.

Let us use *Self* to denote the subjective self, emphasizing that it is from the reflexive consciousness. Self is myself from the standpoint of the first person "I".

Another feature of Self in addition to subjectivity is its *uniqueness*: - Self is distinct from anything else existent in the world. Every person has his/her Self and feels the existence of the world through the Self. Among many consciousnesses related to subjective Self-awareness, there is a key recognition: "I'm alone in this world, - yesterday, today and tomorrow. No one is same as me. If I died, the world around myself would be gone for me forever." Self-awareness enables a person to recognize that nothing or no person is same as his/her subjective Self. S/he is distinct from any other person and anything else in this world.

The distinction between subjective Self and any other objective things can be seen in this way. Think of the answers to the following two questions. Let P_1, P_2, P_3, …denote anything other than subjective Self, where P_i can be of life or of no life. Question 1: To me, the subjective Self, what would this world be like if any P_i is destroyed or dies? Question 2: To me, the subjective Self, what would this world be like if "I" is destroyed or dies? Self's answer to Question 1: The world around "me" would be same as before except that P_i disappears forever, but P_1, P_2, P_3, …, P_{i-1}, P_{i+1}, …are still in the world around me. Self's answer to Question 2: The whole world around me would disappear, including P_1, P_2, P_3, …, and Self, forever. The two answers show the essential difference between Self and non-Self, and the uniqueness of Self.

Although there are multifarious definitions of Self, the above two features, subjectivity and uniqueness, are being accepted in all literatures and among all scholars. Self is in the singular of the first person. No one has ever argued that Self can be in the plural. Objective self's can be plural, such as "themselves" and "yourselves". But subjective self, Self, is always in the singular. It is not possible to have two or more Self's existing at the same time. The essential part of self-awareness or self-consciousness is recognizing the unique specialty of Self: - If Self dies, then the world currently around the subjective "I" will disappear forever.

2.2 An Electronic Robot Can Never Be Self-conscious

Is it possible to have an artificial machine which is self-aware?

At any time point, Self is unique and singular. That means at any time point, it is not possible to have two or more Self's. That is, for an existing Self, it is not possible to have another entity, no matter whether it is nature-made or man-made, which is identical to the Self. The direct logical corollary is: Self cannot be duplicated and copied.

What does "copy" or "duplication" mean? Let us define these common words in more accurately. Object H is a *copy* or *duplication* of object G in terms of J, if they are identical in aspect of J. That is, no one can tell the difference between G and H in aspect of J. Thus, we say G is copiable or duplicatable in terms of J. For example, a document on paper is "copied" on a copy machine. The original and the copy are identical in the aspect of the contents and format, even though they might be different in the other aspects, quality of the paper for example. A computer program for word processing is copiable from a computer to another, because after copying, the codes and functions of the copy are identical with the original, and no one can tell which one is the original and which one is the copy.

Imitating a painting is not duplicating, because at least some top artists can tell the difference between the imitation and the original, even they look same for most of people.

Self cannot be copied in terms of consciousness. Suppose Self S_1 is copied to another entity as S_2. Even though most people cannot tell the difference between S_1 and S_2, at least the original Self S_1 can tell the difference between S_1 and S_2. S_1 would say, "I am still here. S_2 is not myself!" Therefore, S_2 is not a copy of S_1.

Programs of an electronic computer are copiable. A program in a digital computer is a step-by-step procedure or algorithm which can be executed in the computer to accomplish certain function. By the Church-Turing Thesis (Russell and Norvig, 2010) (Turing, 1950), an executable algorithm on a computer can be converted to a set of equivalent 0-1 codes executable on the Turing Machine. Obviously, the 0-1 codes on the tape of the Turing Machine are duplicable or copiable.

Therefore, it is not possible to have an electronic robot to be programmed to have self-consciousness anytime in the future. That is because if there were a robot to be programmed to have subjective Self, then those programs could be duplicated to other robots with the same Self, - which would contradict to the feature of uniqueness of Self.

It is not impossible to have artificial self-consciousness on a man-made machine, but that machine must not be copiable. All the man-made machines currently we have are copiable in terms of the functions. We have not yet had a machine that is not copiable. What an uncopiable machine is like is unknown yet.

3 IMPLICATION AND DISCUSSION

We have logically argued that an electronic robot can never be programmed to be self-aware, therefore they will always lack the so called "self-conscious emotions" (Tracy and Robins, 2004), which are the consciousnesses associated with self-awareness such as shame, pride, self-respect, and self-motivation. An electronic robot therefore will never possess the full range of human consciousness, and will never be a "human". The work on developing self-awareness in electronic computers will end in vain. The researches on the social and legal issues in the future society when robots of full range of human consciousnesses walk all around are based on an unfounded and delusive assumption.

Let us revisit Kursweil's thought lab cited in Section 1. Kurzweil recognized the absurdity occurred between himself and his copy, and should have come to the theory as we derived in Section 2. But he was just stunned by the absurdity, "the copy of me is not me!", with no further probe into "why".

Our reasoning in Section 2 tells the answer to the puzzle in Kurzweil's thought lab: "The so-call 'copy' of Kurzweil is not a copy of Kurzweil in the first place!" Kurzweil's subjective Self cannot be copied.

Now another question comes up: If the "copy" by scanning Kurzweil's brain with reverse engineering is not the copy of his Self, then what is missing in reverse engineering? We do not know. What we know is: the information from scanning the neurotransmitter, synapses, neural connection and every other details of the brain of Self is not sufficient to form Self.

John Searle sensed something wrong with Kurzweil's hypothesis of coping himself by reverse engineering (Searle, 2002), but did not reach the essential of the dilemma either: the Self of Ray Kurzweil cannot be copied.

Can a digital robot be someday as intelligent as, or as spiritual as, a human? This is a long-lasting contentious issue. Wang reasoned that a copiable computer cannot have the consciousness of "fear of death" (Wang, 2013). Our arguments in Section 2 have showed another example of human consciousness, self-consciousness, which cannot be realized in a digital computer. Therefore, a digital computer can never have the full range of human consciousnesses, and will not have souls that are based on self-awareness. Digital robots can never be one of us.

We do not rule out the possibility of having a man-made machine with self-consciousness sometime in the future. But a machine with self-consciousness must be uncopiable in the first place. Conceptually, all the machines that humans have developed are copiable because the hardware of a machine can be copied by reverse engineering, and the software (programs) can be copied per the Church-Turing Thesis. We have not developed a machine which is conceptually uncopiable like Self. We even do not have an idea on what an uncopiable machine is like. The "dream" of having a self-aware humanoid will not come to true soon, even if it will.

Bill Joy once seriously worried about the fate of human beings when computers surpass humans on intelligence. "How soon could such an intelligent robot be built? The coming advances in computing power seem to make it possible by 2030. And once an intelligent robot exists, it is only a small step to a robot species - to an intelligent robot that can make evolved copies of itself." He viewed the research on computer intelligence similar to the research work of atom bombs in 1940's, and called for that "researches leading to the danger should be relinquished." (Joy, 2000). His worry can now be relieved due to the resolution we have derived in Section 2.

Our arguments in Section 2 give a logical answer to the issue everyone many have thought of. The arguments are simple and can be understood by everyone, which are just based on common sense and the fundamentals of logic rules. But why has no one ever logically derived them? People tended to put their opinions based on beliefs, faiths, and subjective judgments, and stay there without going one step further. Some, like John Searle, even asserted that whether a computer may have human consciousness is a problem unable to prove or disprove.

The reasoning addressed in this article is composed of straightforward deductions that everyone is able to do but no one did them. Such a phenomenon is not alone in the history of science. When Stephen Hawking mentioned the big-bang theory of universe, he said, "The discovery that the universe is expanding was one of the great intellectual revolutions of the twentieth century. With hindsight, it is easy to wonder why no one had thought of it before. Newton, and others, should have realized that a static universe would soon start to contract under the influence of gravity. … This behavior of the universe could have been predicted from Newton's theory of gravity at any time in the nineteenth, the eighteenth, or even the late seventeenth centuries. Yet so strong was the belief in a static universe that it persisted into the early twentieth century. Even Einstein, when he formulated the general theory of relativity in 1915, was so sure that the universe had to be static that he modified his theory to make this possible, introducing a so-called cosmological constant into his equations." (Hawking, 1996)

Kurzweil and Minsky recognized that the 'copy' of 'myself' by reverse engineering was not myself. But they did not go one step further for some reason to recognize that the so-called 'copy' is not a copy in the first place. They presumed that all human consciousnesses, including self-awareness, come from conceptually copiable neurons, synapses, molecules and atoms so surely that they would not cast a doubt on that belief even they had come across a logical contradiction. They simply bypassed the logical dilemma.

4 FURTHER RESEARCH

Even though electronic computers will never achieve self-awareness, computers' capabilities of logical deduction and data processing will keep progressing, and computers will achieve some human consciousnesses. Future robots could be very intelligent, very human-like in terms of appearance and action; but they are not self-aware and do not have the consciousnesses related to self-awareness such as shame, pride, self-respect, and self-restraining. What will the world be like by that time? Are we going to treat those humanoids, who are highly intelligent but not self-aware, as machines or as humans?

We need to continue the research on "self-conscious emotions" to identify all self-conscious emotions which can never be achieved on digital robots, so that we can figure out what the future robots are like and better prepare for our future society.

We now have a necessary condition for a machine to be self-aware: the computer must not be copiable. What is an "uncopiable" computer like? How to make such an "uncopiable" computer? These issues are particularly essential for those who are obsessed in developing robots with self-consciousness.

REFERENCES

Hawking, Steven, 1996. *The Illustrated A Brief History of Time*, p.52-53.

Joy, Bill, 2000. Why the Future Doesn't Need Us? Wired. Vol. 9, No.10.

Kurzweil, Ray, 2002. "The Evolution of Mind in the Twenty-First Century," In Richard J. (Ed.). Are we spiritual machine? Discovery Institute Press, Seattle, Washington, p.48.

Kurzweil, Ray, 2005. "The Singularity Is Near – When humans transcend biology," Penguin Books, New York.

Russell, Stuart and Norvig, Peter, 2010. Artificial Intelligence – A modern approach; 3rd edition, Prentice Hall, New Jersey.

Searle, John, 2002. "I Married a Computer", In J. Richards (ed.) Are We Spiritual Machine? – Ray Kurzweil vs. the critics of strong AI. Discovery Institute Press, Seattle, Washington.

Tracy, Jessica L. and Robins, Richard W., 2004, Putting the Self Into Self-Conscious Emotions: A Theoretical Model, Psychological Inquiry, Vol. 15, No. 2, 103–125.

Turing, Alan, 1950. "Computing machinery and intelligence," *Mind*, Vol.59, 433-466.

Wang, Jinchang, 2013. "On the Limit of Machine Intelligence," *International Journal of Intelligence Science*, Vol. 3, No. 4, 170-175.

Wikipedia 2014 (1) http://en.wikipedia.org/wiki/Self.

Wikipedia 2014 (2) http://en.wikipedia.org/wiki/Self-awareness.