Artificial Qualia, Intentional Systems and Machine Consciousness

Robert James Boyles
Department of Philosophy, De La Salle University-Manila
2401 Taft Avenue, Manila, Philippines

In the field of machine consciousness, it has been argued that in order to build human-like conscious machines, we must first have a computational model of qualia. To this end, some have proposed a framework that supports qualia in machines by implementing a model with three computational areas (i.e., the subconceptual, conceptual, and linguistic areas). These abstract mechanisms purportedly enable the assessment of artificial qualia. However, several critics of the machine consciousness project dispute this possibility. For instance, Searle, in his Chinese room objection, argues that however sophisticated a computational system is, it can never exhibit intentionality; thus, would also fail to exhibit consciousness or any of its varieties. This paper argues that the proposed architecture mentioned above answers the problem posed by Searle, at least in part. Specifically, it argues that we could reformulate Searle's worries in the Chinese room in terms of the three-stage artificial qualia model. And by doing so, we could see that the person doing all the translations in the room could realize the three areas in the proposed framework. Consequently, this demonstrates the actualization of self-consciousness in machines.

1. INTRODUCTION

The field of machine consciousness (MC) focuses on developing machines that both have an inner world (i.e., subjectivity) and artificial consciousness (AC). Several critics, however, dispute the realizability of this project. Addressing the objections posed by Searle (1980) against computational systems, this paper argues for the plausibility of self-consciousness in machines.

2. CONSCIOUSNESS AND AI

Consciousness is often deemed integral in discussing intelligence and the mental lives of agents (Pfeifer and Scheier, 1999). In MC, modeling and implementing artificial consciousness in machines are at the top of the list of concerns. On the one hand, a number of researches in this field deal with subjectivity and consciousness architectures for machines. Several others, on the other hand, focus on using machine models for studying consciousness in general. Holland (2003) distinguishes between the two by differentiating strong AC from weak AC.

Synonymous to the distinction between weak and strong artificial intelligence[1] (AI), Holland states that strong AC is geared towards building conscious machines. Meanwhile, he defines weak AC as designing machines that merely simulate consciousness. It has been argued that the differences between the two, nevertheless, are blurred and quite inseparable in practice[2] (Clowes et al., 2007). And, at present, findings in MC have inspired current AI research that the task of building human-like conscious machines seems to be closer than ever. Yet, many dispute the possibility of implementing consciousness in machines. Searle, for instance, has posed objections against the strong AI thesis; thus, also entailing his unfavorable stance towards the machine consciousness project.

3. STRONG AI TO STRONG AC

Since its inception, there have been many criticisms against the strong AI project like the ones given by Dreyfus (1979, 1992), who has specifically argued against classical artificial intelligence. For his part, Searle (1980) provides his own critique through his Chinese room argument, which supposedly demonstrates that all computational systems fail to exhibit intentionality.

In the thought experiment, Searle asks us to imagine a man (i.e., Searle himself) locked inside a room. The man trapped inside this room does not know any Chinese to a point that, for him, Chinese symbols are just meaningless squiggles. Now, suppose that the man is given a first batch of Chinese writing through slipping pieces of papers inside the room. Afterwards, he is again given a second batch of Chinese writing but, along with this, he also receives a set of rules for correlating it with the first batch. The set of rules are in English, and the man just so happens to be a native English speaker. Note here that the only way he can correlate the two batches is by identifying the different Chinese symbols based on their shapes. Then, he is given a third batch of Chinese writing and another set of rules in English on how to correlate the last batch with first two batches.

Searle further explains that the first batch of Chinese writing is actually a script, the second batch a story, and the last one are the questions. Meanwhile, the set of rules is equivalent to a computer program, which enables the man to answer questions about the story through manipulating Chinese symbols based solely on their shapes. Thus, he would be able to answer certain questions in Chinese even without understanding the language itself. Searle then asks us to imagine that, after a while, the man becomes really good (i.e., efficient) at answering questions in Chinese through memorizing the

[1] As defined by Searle (1980), weak AI is the position that claims that computers are useful tools in studying the mind. Strong AI, on the other hand, claims that the "appropriately programmed computer really is a mind."

[2] Likewise, Chella and Manzotti (2007) pose the question: "For instance, if a machine could exhibit all behaviors normally associated with a conscious being, would it be a conscious machine?"

ISSN 2012-3477

rules. From the point of view of someone asking the questions, the man does understand Chinese given that he can supply correct answers satisfactorily. However, Searle maintains that this is not the case.

Using this Chinese room objection, Searle argues that it can never be said that computer programs are actually thinking. As he points out in his thought experiment, even if the man inside the room knows how to answer Chinese questions in Chinese, he actually does not understand anything at all beyond the manipulation of meaningless squiggles. To quote Searle, he states that (1980):

> "Intentionality in human beings (and animals) is a product of the causal features of the brain. Instantiating a computer program is never by itself a sufficient condition for intentionality... The form of the [Chinese room] argument is to show how a human agent could instantiate the program and still not have the relevant intentionality… Any attempt to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain."

In short, intentionality, for Searle, cannot be realized by any computer program. And, it is not only intentionality that is at stake in Searle's thought experiment. Many have argued that, although the Chinese room was originally put forward by Searle as a direct objection against intentionality, it can also be seen as an argument against consciousness—that is, no computer program can exhibit consciousness or any of its varieties (e.g., self-consciousness, introspection, reportability).

Chalmers (1997), for instance, maintains that, if Searle's thought experiment succeeds, it establishes that the Chinese room system also fails to realize consciousness. Thus, it could be said that consciousness is actually at the center of the Chinese room argument.

Now, in order to show that consciousness could be implemented in machines, addressing the problems posed by Searle is a necessary step. For MC to overcome such objections, Chalmers best puts it that consciousness should taken *seriously.* Capture consciousness, and we have captured intentionality as well. And in the field of machine consciousness, models that could supposedly generate consciousness have been proposed including self-consciousness architectures.

4. ROBOT ARTIFICIAL QUALE

As mentioned earlier, it seems like that once we have already accounted for consciousness, it follows that we have also taken intentionality into account. In MC, architectures such as the ones presented by Chella and Gaglio (2008) purportedly support qualia in machines.

Focusing on how to model robots with self-consciousness, Chella and Gaglio maintain that robot artificial quale is realizable by implementing a robot cognitive architecture with three computational areas, namely: the (1) subconceptual, (2) conceptual, and (3) linguistic areas. Brief descriptions of these areas are as follows:

1. *Subconceptual area:* Devoted to processing all the information coming from the sensors of a robot
2. *Conceptual area:* Mediates the processes of the subconceptual and linguistic areas
3. *Linguistic area:* Area where linguistic representation occurs, which somehow corresponds to a robot's long-term memory

Chella and Gaglio argue that these computational areas generate artificial qualia in machines. They claim that a robot with such architecture is capable of processing information from its built-in sensors (i.e., through its subconceptual area) with the 3D information stored in its conceptual area. Thus, the process leads to the generation a 2D viewer-dependent reconstructed image of a scene that the robot currently perceives.

The reconstructed 2D model is not static image in the sense that its construction is done with an active process. Chella and Gaglio further maintain that the bases of the robot's artificial qualia are the conceptual and linguistic areas.

In short, Chella and Gaglio are presenting a model that possibly gives robots the capability of self-consciousness (Chella and Manzotti, 2007). In this proposed architecture, the higher-order of perception of the robot is the basis of its self-consciousness. It can then be argued that, if indeed the model supports self-consciousness, the agent itself is in fact conscious. Note here that self-consciousness has been considered a type of consciousness in general. And it has been argued by some (Chalmers, 1997) that this kind of consciousness, along with the other varieties, has both phenomenal and psychological aspects.

Assuming that self-consciousness could indeed be modeled in machines, we could then reformulate Searle's Chinese room system in terms of Chella and Gaglio's three-stage artificial qualia model.

5. AC REFORMULATION

To re-cast the Chinese room system so that it too exhibits self-consciousness (i.e., including intentionality), let us implement to this the robot cognitive architecture that employs the three computational areas proposed by Chella and Gaglio[3]. And by doing this, it could be seen that the man inside the room performing all the translations, if further argued, could realize the three areas in the proposed framework.

[3] A colleague, Jeremiah Joven Joaquin, has mentioned that what is actually being done here is to present a modified version of the systems reply to Searle's Chinese room argument.

ISSN 2012-3477

In the Chinese room system, it can be said that Chella and Gaglio's subconceptual area has already been accounted for. Recall that this area is responsible for processing all the information from the sensors of a robot; thus, it could be argued that the man inside the room, who in the process of performing all the translations, is already implementing the subconceptual area. The next step is then to implement the conceptual and linguistic areas.

After gathering the relevant information (e.g., the presence of a Chinese symbol **算**$_7$ in front of him), let us try to implement the conceptual and linguistic areas by adding two more men inside the room. Suppose that, at time $\mathbf{T}_1$, the first man **X** is continuously processing **算**$_7$, and he simultaneously sends this information to a second man **Y**. **Y**'s task then is to solely process **算**$_7$ and correlate it with a set of Chinese symbols that was originally given, or "programmed," to him. This set contains different types of the Chinese symbols (**算**, **台**, **叫**…), which also stores the different tokens of these symbols (**算**$_1$, **算**$_2$, **算**$_3$… **台**$_1$, **台**$_2$, **台**$_3$… **叫**$_1$, **叫**$_2$, **叫**$_3$…) that the system has previously encountered. The next step for **Y** is to find a match between **算**$_7$ and the symbol that closely resembles it. For instance, let **算**$_6$ be the closest Chinese character that resembles **算**$_7$. Further, **Y** would then have to generate a viewer-dependent reconstruction that resembles **算**$_7$ (i.e., by constantly matching the symbol **算**$_7$ with **算**$_6$), while simultaneously correlating it with the sensory data being currently viewed by **X** at $\mathbf{T}_2$, or any succeeding time after $\mathbf{T}_1$. As for the third man **Z**, he is now then capable of generating and fixing linguistic representations to the viewer-dependent reconstruction of **算**$_7$.

In this reformulation of the Chinese room, it can be said that the modified system now supports self-consciousness wherein three men are just performing tasks that were just programmed to them. However, it can be further argued that, given that these men are just executing preset operations, it also seems possible to implement the said programs into three distinct robots, $\mathbf{X}_r$, $\mathbf{Y}_r$, and $\mathbf{Z}_r$. Now it is quite reasonable to think that these robots would accomplish their specified tasks as effectively as their human counterparts. Finally, given that it seems possible to design a single program that could execute the tasks identified in the subconceptual, conceptual, and linguistic areas, why not just develop an analogous architecture for a single robot? It can then be argued that this is what Chella and Gaglio have in mind in their three-stage artificial qualia model. The Chinese room system reformulated so that it now supports a variety of consciousness, self-consciousness.

6. CONCLUSION

It seems like that Searle's' Chinese room system could be reformulated in terms of Chella and Gaglio's three-stage artificial qualia model. After modifying the system, it could be argued that the person doing all the translations realizes the three computational areas proposed in the framework. Thus, this demonstrates the actualization of self-consciousness in machines.

7. REFERENCES

Chalmers, D. (1997) *The conscious mind: In search of a fundamental theory*, Oxford University Press, New York.

Chella, A. and Gaglio, S. (2008) 'In search of computational correlates of artificial qualia' in *AGI 2009: Proceedings of the Second Conference on Artificial General Intelligence*, Atlantis Press, Arlington, Virginia, pp. 13–18.

Chella, A. and Manzotti, R. (2007) *Artificial Intelligence and Consciousness*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.5545&rep=rep1&type=pdf (Accessed 21 January 2012).

Clowes, R., Torrance, S. and Chrisley, R. (2007) 'Machine consciousness: Embodiment and imagination', *Journal of Consciousness Studies*, Vol. 14, No. 7, pp. 7–14.

Dreyfus, H.L. (1979) *What computers can't do: A critique of artificial reason*, The MIT Press, Cambridge, Massachusetts.

Dreyfus, H.L. (1992) *What computers still can't do: A critique of artificial reason*, The MIT Press, Cambridge, Massachusetts.

Holland, O. (Ed.), (2003) *Machine Consciousness*. Imprint Academic, New York.

Pfeifer, R. and Scheier, C. (1999) *Understanding Intelligence*, The MIT Press, Cambridge, Massachusetts.

Searle, J.R. (1980) 'Minds, brains, and programs', *Behavioral and Brain Sciences*, Vol. 3, No. 3, pp. 417–457.