Proposal for an Approach to Artificial Consciousness Based on Self-Consciousness.

Christophe Menant

No Affiliation
75 R. G. Mandel - 33000 Bordeaux – France -
christophe.menant@hotmail.fr

Abstract

Current research on artificial consciousness is focused on phenomenal consciousness and on functional consciousness. We propose to shift the focus to self-consciousness in order to open new areas of investigation. We use an existing scenario where self-consciousness is considered as the result of an evolution of representations. Application of the scenario to the possible build up of a conscious robot also introduces questions relative to emotions in robots. Areas of investigation are proposed as a continuation of this approach.

Artificial Consciousness and Consciousness

Current research on artificial consciousness relies on phenomenal consciousness as a reference model for human consciousness, or introduces functional consciousness to cover the functional aspects of human consciousness. We would like to propose here an alternate approach that takes self-consciousness as a reference model. Such approach can open new areas of investigation for artificial consciousness and bring new items for research on human consciousness.

Self-consciousness can be defined as “the possession of the concept of the self and the ability to use this concept in thinking about oneself” (Block, 2002). It is different from phenomenal consciousness which can be understood as “experience; the phenomenally conscious aspect of a state is what it is like to be in that state” (Block, 2002). We feel that taking self-consciousness as a reference for analyzing the possibilities of an artificial consciousness can open new perspectives for robots perception, reasoning and action. Such approach can also make available a new frame for positioning artificial consciousness versus human consciousness, understanding that the relations between phenomenal consciousness and self consciousness still remain to be clarified.

Self-Consciousness and the Evolution of Representations

Most people agree that human consciousness is a product of evolution. We believe that human self-consciousness can be considered as the result of an evolution of representations that brought animal life from a non self-conscious auto-representation to the level of a conscious self-representation.

A scenario has already been proposed to cover such an evolution of human self-consciousness (Menant, 2006). We would like here to see how that scenario based on the evolution of representations can be applied to artificial consciousness. The core of our approach is the evolution of representations in organisms, up to the level of a conscious self-representation that introduces self-consciousness. The human scenario can be summarized in four steps:

A) Early primates had representations of their con-specifics as organisms existing in the environment, and had also an "auto-representation" (representation of himself for a subject, but devoid of any notion of self-consciousness).

B) Pre-human primates were capable of a level of inter-subjectivity (Gardenfors, 2006) that gave them the possibility of some identification with their con-specifics, like our today great apes. We consider that this performance allowed the auto-representation of a subject to merge with the representations he had of his con-specifics.

C) Such merger of representations brought the auto-representation to access the meanings associated to the representations of the con-specifics, namely the meanings corresponding to organisms perceived as existing in the environment with an identity, and recognized as such by the subject and by his con-specifics. We consider that these new meanings allowed an evolution of the auto-representation of the subject into a beginning of conscious self-representation.

D) The performances associated to these first elements of conscious self-representation brought up some evolutionary advantages that allowed its evolution into self-consciousness. These evolutionary advantages are closely related to the rooting of self-consciousness in emotions, as there is some rational to consider anxiety as an evolutionary engine in the scenario. Anxiety can be generated or amplified by the identification with endangered or suffering con-specifics (this happened at pre-human primate time frame where the survival of the fittest was the law). Too much anxiety is difficult to live with and has to be limited. We propose that anxiety limitation for pre-human primates lead to the development

of empathy, imitation, language and inter-subjectivity, which in turn created a positive feedback loop on the identification with con-specifics. This anxiety limitation process positions anxiety as an evolutionary engine in human evolution. It can also provide a research thread for a phylogenesis of human emotions.

Tentative Application to Artificial Consciousness

We propose an application of the scenario to artificial consciousness with a robot going thru the four steps. Step A) needs the build up of a robot that carries a non self-conscious auto-representation and has representations of his con-specifics. This looks possible in principle, as a continuation of what has been initiated in terms of representation and meaningful information (Menant, 2005). Step B) is about the robot merging his auto-representation with the representations he has of his con-specifics. As a parallel with human evolution, this performance could come up thru some kind of inter-agentivity between robots where a robot becomes capable of understanding or guessing what a con-specific robot is to implement. This needs some clarification of the relations between "theory-theory" and "simulation-theory" used in theory of mind (Gordon, R. 2004). This step also needs an understanding of the data processing content associated to a merger of representations. Some conceptual effort may be necessary for the implementation of this step in a robot. Step C) is a key step where the auto-representation gets access to the meanings associated to the representations of the con-specifics. More precisely, the auto-representation of the robot gets access to the meaning of "entity existing in the environment". At this level, we would have the first elements of a conscious self-representation appearing in a robot by his auto-representation becoming an existing element in his own environment. This looks like the toughest part to analyze and realize. We would use the notions introduced about representations and meaningful information in order to analyze the data processing related to the merger of representations that explicitly includes the meanings.

Regarding step D), the evolution of the robot from a limited conscious self-representation to self-consciousness may or may not use the same evolutionary engines that the ones we assume were active during human evolution. The anxiety limitation process is not mandatory for the evolution of robots. We can program robots with other evolutionary engines. It could be decided to favorize empathy between robots (for their group survival), as well as imitation, language and inter-agentivity, but there is no obligation to relate them with anxiety limitation. Such choices will condition the content of the emotional world of the robots resulting from this evolutionary process. The emotional world of a self-conscious robot can be close to our human one or be very different depending upon our choices. But, whatever the choices, ethical concerns will appear.

Such approach to artificial consciousness as based on self-consciousness also naturally introduces the questions related to machine autonomy and free will. These subjects are beyond the scope of this paper.

Conclusion and Continuation

We have looked at the possible build up of a self-conscious robot using a human evolutionary scenario based on an evolution of representations. Our starting point is a non self-conscious auto-representation of a robot. Robots group life is a key point in the scenario as the auto-representation is to merge with the representations of con-specifics robots, allowing the auto-representation to become "existing in the environment". The resulting conscious self-representation of the robot introduces self-consciousness for robots as a new entry point to artificial consciousness.

Continuation of the proposed approach will include the following:

- Complement the definition of a representation as based on meaningful information. Define the content of a merger of representations, including their associated meanings.
- Analyze the definitions and the contents of possible evolutionary engines as specific to self-conscious robots.
- Propose an evolutionary nature of the self, which would include first and third person components, by using the identification of the auto-representation of the subject with the representations he has about his con-specifics.
- Investigate the nature of emotions, free will and autonomy for robots as part of the proposed evolutionary scenario. Consider the corresponding consequences as part of the moral dimensions of artificial consciousness.
- Look at how this evolutionary scenario could help finding relations between phenomenal and self-consciousness.

References

Block, N. 2002. Some Concepts of Consciousness. In *Philosophy of Mind: Classical and Contemporary Readings*, David Chalmers (ed.) Oxford University Press.

Gardenfors, P, 2006. On the Evolution of Intersubjectivity. To appear in *Consciousness Transitions - Phylogenetic, Ontogenetic and Physiological Aspects*, ed. by H Liljenström and P. Århem, Elsevier.

Gordon, R. 2004. Folk Psychology as Mental Simulation. *Stanford Encyclopedia of Philosophy*.

Menant, C. 2005. Information and Meaning in Life, Humans and Robots. In *Proceedings of FIS2005.*

Menant, C. 2006. Evolution of Representations and Inter-Subjectivity as Source of the Self. An introduction to the Nature of Self-Consciousness. In *Proceedings of the 10th ASSC.*