ELSEVIER

A philosophical assessment of computational models of consciousness

Action editor: Ron Sun

Selvi Elif Gök [a,*], Erdinç Sayan [b]

[a] *Informatics Institute, METU, İnönü Bulvarı, 06800 Ankara, Turkey*
[b] *Department of Philosophy, METU, İnönü Bulvarı, 06800 Ankara, Turkey*

Abstract

There has been a recent flurry of activity in consciousness research. Although an operational definition of consciousness has not yet been developed, philosophy has come to identify a set of features and aspects that are thought to be associated with the various elements of consciousness. On the other hand, there have been several recent attempts to develop computational models of consciousness that are claimed to capture or illustrate one or more aspects of consciousness. As a plausible substitute to evaluating how well the current computational models model consciousness, this study examines how the current computational models fare in modeling those aspects and features of consciousness identified by philosophy. Following a review of the literature on the philosophy of consciousness, this study constructs a list of features and aspects that would be expected in any successful model of consciousness. The study then evaluates, from the viewpoint of that list, some of the current self-claimed and implemented computational models of consciousness. The computational models studied are evaluated with respect to each identified aspect and feature of consciousness.
© 2011 Elsevier B.V. All rights reserved.

Keywords: Consciousness; Computational cognitive modeling; Clarion; LIDA; ACT-R; Neuronal Work Space Model; ART; GMU-BICA

1. Introduction

There has been a recent flurry of activity in consciousness research. Consciousness is an inherently difficult subject both in terms of philosophical understanding and in terms of pragmatic results that could be used in engineering applications. Notwithstanding the fact that there is no framework at the present for studying consciousness that enjoys universal acceptance, philosophy has come to identify a set of features and aspects that are thought to be associated with the various elements of consciousness. On the other hand, there have been several recent attempts to develop computational models that are claimed to capture or illustrate one or more aspects of consciousness. Being a very complicated issue, there are many alternative views of consciousness around. This study takes as its departing point some of the major viewpoints currently available in philosophy. We first study the various features and aspects of consciousness that can be found in the literature on the philosophy of consciousness. We then examine some self-claimed[1] computer models of consciousness and evaluate these models according to how well they accommodate and explain the various features and aspects of consciousness pointed out by philosophers. Moreover, we restrict our study to those models that have been implemented and whose behavior studied. The complete results of this evaluation and survey not only rank the models according to their proficiency, but also present general clues as to how successful cognitive science currently is when it comes to the scientific understanding of consciousness. In this respect, this study combines philosophy and computer

---

* Corresponding author. Tel.: +90 312 210 38 09.
E-mail address: elifg@ii.metu.edu.tr (S.E. Gök).

[1] The reason we call them "self-claimed" is to indicate the point that the developers of these computer models claim that these models can capture certain aspects of consciousness.

science in reviewing recent work in consciousness research in both fields.

In evaluating computational models of consciousness from the viewpoint of philosophy, this study hopes to have achieved three things. First, it has identified and systematically compiled those aspects and features of consciousness that have appeared in recent philosophy of mind literature. As we took the union of all of the features and aspects, some inconsistencies in approach had to be resolved. The resultant eclectic and consolidated list of philosophical features and aspects of consciousness is a product of this work. Second, it examined the current set of self-claimed computational models of consciousness vis-à-vis our list. The examination reveals to what extent the computational models satisfy the list, and hence, the concerns of philosophy. It can be argued that any successful model must, at a minimum, satisfy our eclectic list to be acceptable by the philosophy community. Third, the work provides an extensible framework of organization and structure that could aid and help to direct future efforts in the interdisciplinary fields of consciousness. Being from diverse conceptual backgrounds, a unifying framework can act as a facilitator or mental aid in mediating the different views of researchers that make up the constituents of the interdisciplinary area of study. Take, for instance, the case of phenomenal consciousness, whose explanation is considered by some philosophers to be unattainable by scientific means, at least at the present state of knowledge. Some computational models do attempt to handle phenomenal consciousness. If, in some future study, computational models provide a higher level of scientific understanding of phenomenal consciousness, then such discovery would be of paramount importance to philosophers in revising their apprehension of the issue of consciousness. Such revisions may even have a cascading effect, whereby other derivative concepts of philosophy would ultimately benefit from the computational model. All this, however, requires that the interdisciplinary approach is embedded in a cohesive operational framework, understood and used by all parties.

It should be noted that this study has a particular weakness. Our evaluation of computer models is entirely based on the literature about the models. That is, we investigate the mechanisms of a particular model through the literature devoted to the model. We did not do any hands-on work on the models. Although this is a particular weakness of the study, it was necessitated by certain practical reasons. Firstly, reaching the source code of some models, like IDA that is developed for the US Navy, is not possible. Also, one needs to be competent in both the languages and the environments of the computer models in order to perform hands-on work at a level sufficient to fully evaluate the models. Otherwise, the reason for a possible failure in modeling a particular task will be open to debate, about whether the model itself is incapable of the task, or the user has insufficient knowledge to implement the task. When these two concerns are taken into consideration, we think that it is justifiable that this study limits itself to the

descriptions and claims in the literature, and does not complement it with practical work.

2. The elements of consciousness

Before we present our composite and eclectic list of features of consciousness, we first examine the various such lists from several researchers. However, most of the features in these previous lists have direct references to the philosophy of consciousness literature without further explanation. A full appreciation of these lists requires a brief review of the current state of the art in philosophy of consciousness.

The body–mind problem is an ancient one that has an important place in philosophy. However, consciousness can be seen as a relatively new concept as it is used in the current philosophy and cognitive science literature. The usage of the term 'consciousness' can be traced back to Descartes. He used the term to refer to the inner knowledge of the subject. Descartes (1973, p. 222) also defined thought as "all that of which we are conscious as operating in us." It should be noted that from Descartes until very recently, consciousness was taken as the essential characteristic of the mental. That is, it was thought that there were no such things as unconscious mental states. However, as an exception, Leibniz (1989, pp. 295–299) made a distinction between what he called "petit perception" and "apperception." Petit perceptions are the perceptions that the subject is not aware. The combination of petit perceptions leads to apperception. Apperception can be seen as perceiving that is also accompanied by an awareness of perceiving. Yet, the state of art equated consciousness with the totality of mental states until Freud. Freud can be considered as the first who conceptualized an elaborate framework for unconscious mental states.[2]

By the early 20th century, the field of psychology had seen the rise of behaviorism. According to Baars (1986) behaviorism is a metatheory of psychology and each metatheory "specifies a domain for psychology, a set of techniques for investigating that domain, and a research program to integrate the findings into the body of human knowledge and practice" (p. 5). Behaviorism rejected introspection to be used as a part of methodology in psychology, and proposed that the only proper domain of psychology is the observable human behavior. So, with the rise of behaviorism, not only consciousness but also any kind of investigations concerning the hidden nature of mental states had been left out of

---

[2] It should be noted that the "Freudian unconscious" is different from what one may call the "cognitive unconscious." There are two kinds of "unconscious" in the Freudian framework. The term 'unconscious proper' stands for the mental states or processes that were conscious for some time but are now repressed. The unconscious proper can be made conscious through psychoanalysis. On the other hand, there are preconscious mental states or processes that are only temporarily unconscious and can become conscious without any special technique. Whereas the "cognitive unconscious" indicates the processes that underlie cognition that are not and cannot become conscious (Güzeldere, 1997, pp. 20–21).