Available online at www.sciencedirect.com

SciVerse ScienceDirect

Cognitive Systems Research 17–18 (2012) 49–62

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

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*

Received 20 July 2011; accepted 19 October 2011
Available online 20 November 2011

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.

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.

1389-0417/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2011.11.001

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).

science. However, by the cognitive turn that took place in the middle of the 20th century many mental processes took their place back in psychology. The claim of cognitive psychologists is that "psychologists observe behavior in order to make inferences about underlying factors that can explain the behavior" (Baars, 1986, p. 7). Yet, consciousness was still avoided as a subject of scientific investigation until recently. The studies of consciousness became popular in the 1980s when empirical findings on unconscious processes started to accumulate. This accumulation gave way to treating consciousness as a variable[3] and studying it empirically (McGovern & Baars, 2007).

Correspondingly, there has been a rise of consciousness studies in the field of philosophy also. Besides several philosophical theories that propose some explanation about how mental states become conscious, there is also an extensive literature that is devoted to the conceptual clarification of the issue. In addition to general philosophical theories which try to locate the place of mind and of consciousness in the entire reality (as part of the project of solving the body–mind problem), there have been proposed, in the last few decades, theories more specifically focusing on the phenomenon of consciousness and its various aspects. Theories of the former kind, i.e. theories of general metaphysical nature, can be grouped under two broad headings: physicalist theories and non-physicalist (mostly dualist) theories. A classic example of a dualist theory is Descartes' substance dualism. While the long history of the non-physicalist theories goes back to at least Plato, approaches to the body–mind problem in more recent times have been predominantly physicalistic. Theories of the latter kind, i.e. contemporary theories that specifically focus on the phenomenon of consciousness, tend to be decidedly physicalistic. Those theories can be grouped under several classes. The "higher-order theories" claim that consciousness arises when our experiences are attended by mental states that are about those experiences. If the higher-order state that is about the lower-order states is what may be characterized as a thought, the account is a "higher-order thought (HOT) theory" (e.g. Rosenthal, 1997); but if the higher-order state in question is taken to be akin to perception, then it is more properly labeled as a "higher-order perception (HOP) theory" (e.g. Armstrong, 1997; Lycan, 1997). The representational theories, by contrast, maintain that a mental state's being conscious consists of nothing but the entire representational content of that mental state (e.g. Tye, 2000). Also available on the market are theories offering to account for consciousness in terms of the notion of "global workspace" (Baars, 1988) (we will talk about this theory in more detail later in connection with LIDA), or of "integrated information theory" (Tonini, 2004). There are also more exotic accounts resorting to quantum phenomena allegedly taking place in certain structures inside the neurons called *microtubules* to illuminate the mysteries of consciousness (Hameroff & Penrose, 1995).

The theories of consciousness all attempt to solve what they take to be the problem of consciousness or its various facets. However, it is not clear whether there really is a "the problem of consciousness," as it seems that we do not always talk about the same concept when we talk about consciousness. As Block puts it, "The concept of consciousness is a hybrid or better, a mongrel concept: the word 'consciousness' connotes a number of different concepts and denotes a number of different phenomena" (Block, 1997, p. 376).

Moreover, it has also been argued that there are no analogs of consciousness in nature due to its subjectiveness. One of the best-known arguments concerning the limits to any possible scientific explanation of consciousness is that proposed by Nagel (1997) in his famous article "What is It Like to be a Bat?". In this article, Nagel says "the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism" (1997, p. 519). Accordingly, any explanation of consciousness must also be able to account for 'what it is like to be' such a conscious organism. However, if one tries to understand this aspect, the only thing one may be able to do is to imagine what it is like for us to *behave like* such an organism. That is, one will be always missing the crucial part, i.e. the point of view of the organism itself. After all, to have a conscious experience means having a subjective point of view. It is this point that puts consciousness in a different context than all the subject matters of physics. Physics handles its subject matter objectively – at least it should do so. Hence, as Nagel (1997, p. 524) puts it, "If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done." Accordingly, the subjective experience, hence consciousness, does not seem to be explicable in any physical theory, which by its nature shuns making room for subjectivity and instead tries to take an objective viewpoint of the world.

A similar argument is the "Knowledge Argument" proposed by Jackson. In this thought experiment, one is invited to consider the situation of Mary (Jackson, 1997, pp. 567–570). Mary is a brilliant scientist, who grew up in a black-and-white room. Nevertheless, she knows everything there is to be known to physics, which she learned from her black-and-white television and books. The question is whether or not Mary knows everything there is to know about sensations of color. Jackson argues that she cannot. He considers what will happen if Mary is released from her room and sees the color red for the first time. Jackson's argument is that at that moment Mary learns something new: she learns the redness of red or what it is like to see red.

[3] One can treat consciousness as a variable in the sense that there may be empirical studies focusing on the mental states or the mental processes of the subjects where one group of subjects are conscious of their mental states or processes, in contrast with another group who are not conscious of them.

Levine (1997) accepts that Mary learns something new by experiencing red color. However, what she learns is not necessarily a nonphysical fact. But rather, it is a fact that is not explicable in terms of scientific propositions. Accordingly, he announces the existence of an "explanatory gap" between the explanatory capabilities of physical theories that are proposed in the scope of science and what is presented in a conscious experience, such as the subjective character of a sensation of red color.

2.1. Lists of features of consciousness

Each of us knows something about consciousness. We have some intuitions about it. In many other domains, one can keep theorizing without taking our intuitions into consideration. However, in the case of consciousness, it seems that we all want a theory that gives us an explanation about our own intuitions. This is not to say that, once the correct theory of consciousness is formulated, all our intuitions will turn out to be scientific truths – they may be illusions. However, any such theory should at least give us an understanding of the mechanisms underlying these illusions.

Given the above considerations, it is quite clear that we do not have an operational definition of consciousness. Even the possibility of such a definition is questionable. In such a situation, it seems plausible to start with a conceptual analysis in an attempt to gain a foothold on the issue.

There are some studies in philosophy that try to identify the various features and aspects of consciousness. One of the earlier analyses of the features and aspects of consciousness is due to Block (1997). He differentiates four kinds of consciousness: phenomenal consciousness, access consciousness, monitoring consciousness, and self-consciousness. Block regards his analysis as one partly aiming at regimenting our concept of and terminology about consciousness. That is, the four types of consciousness that he identifies are not necessarily four different kinds of consciousness in existence. It may turn out that when consciousness studies further evolve, some or even all of these types are reduced to a single form. However, when the current state of the art is considered, the classification seems, at least at a conceptual level, both necessary and useful. It is necessary in the sense that if any theory of consciousness claims to give an exhaustive explanation of the issue, it should give account to either all four types, or to the reduced forms. Also, if we do not distinguish different kinds of consciousness, there is a danger that we end up with a theory that explains only one kind.

Van Gulick (1995) proposes six features of consciousness that must be explained if we want to understand consciousness. His main concern is about the clarification of the subject matter of consciousness studies. He states that before we start talking about the possibility of studying consciousness scientifically, we must be clear about what to expect from such a study. So, the list that he proposes can be taken as consisting of the features of consciousness for which we expect an explanation from a scientific study of consciousness.

The first thing that a scientific study of consciousness must explain is the difference between conscious, unconscious, and nonconscious mental states. Another distinction is the difference between conscious and unconscious creatures. Creature consciousness may be seen as a property that is ascribed to some creatures.[4] It is the creature consciousness that one talks about when stating that an amoeba may not be conscious at all. According to Van Gulick, our being able to use its content directly is yet another feature of consciousness. That is, we have direct knowledge about the content of a conscious mental state. As Van Gulick (1995, p. 65) puts it, conscious mental states "have meaning or content for the person or creature whose states they are." The last three features in the list – namely qualia, phenomenal structure and subjectivity – have a common point. Van Gulick states that these three features are usually referred to as the phenomenal aspects of consciousness. However, it is better for us to distinguish these three, since each seems to carry a special essence that is not covered by the others.

Lycan (1999) points to the confusion in the literature on consciousness. He, like Block, states that there are some philosophers and scientists who seem to confuse the different problems of consciousness. Lycan states that distinguishing these different problems would not only prevent confusion, but would also help us to advance further in consciousness studies. Different problems of consciousness that Lycan identifies correspond to the different features of consciousness.

The first problem that Lycan mentions is the difference between conscious and unconscious states. Secondly, one seems to have knowledge of the content of one's conscious experiences via introspection. This introspective knowledge is not readily available to any other person. Another problem has to do with the concept of qualia. The concept of qualia takes its place among the problems that must be investigated separately, according to Lycan. Conscious experiences have smoothness and contiguousness, which Lycan calls homogeneity, that seem to be lacking in the external physical world. Despite the discrete nature of the properties of physical materials, our experience of these properties is smooth and continuous. This homogeneity must also be explained. Yet another feature is the intrinsic perspectivalness of conscious experiences, i.e. the first-person perspective of the conscious experiences. This perspective is the subjective point of view that Nagel (1997) emphasizes as a leading obstacle to the studies of consciousness.

[4] We can distinguish two senses of consciousness: *creature consciousness* and *state consciousness*. Creature consciousness denotes a creature's property of being conscious. On the other hand, state consciousness is a property of mental states: a creature's mental states can be conscious or unconscious.

The other three problems are the "funny facts," the ineffability of "what it's like," and the explanatory gap, as Lycan calls them. He states that the knowledge argument (Jackson, 1997) reveals that there may be some facts that are nonphysical. These facts, which Lycan calls "funny facts," need an explanation. One seems to be incapable of explaining what it is like to have a particular conscious experience, which stands out as another problem. Lastly, Lycan points to the "explanatory gap" problem that Levine (1997) raises. Lycan states that the connection between the subjective properties of a conscious experience and physical, i.e. neurological, facts must be somehow explained.

2.2. An eclectic and consolidated list of the features of consciousness

Our list, which is presented below in Table 1, is derived on the basis of the three authors' lists that were examined in the previous section. Most of the features of consciousness that are identified in the different lists we mentioned resemble one another. So, it should suffice to include the similar features that appear in different lists as a single entry in our list. Secondly, we omit some of the features given by some of the authors.

The order of the items in the list is not meant to indicate any hierarchy. However, the first item, namely Difference (between conscious and unconscious mental states) has a special place with respect to any theory of consciousness, as we shall see in a moment. Also, the last three items, i.e. Qualia, Phenomenal Connectedness, and Subjectivity, share relevance with phenomenal consciousness. As such, they appear as contiguous items in Table 1. Apart from these, there is no further meaning to the order of the items.

The first element in our list concerns the difference between conscious and unconscious mental states. That is, a theory of consciousness must explain how some mental states become conscious and others do not. This item is different from all the others in the sense that it is the most obvious aspect to be explained. In fact, this first element could be a good starting point in the establishment of any comprehensive theory that attempts to explain consciousness. This item appears in two of the lists that we considered above, namely the lists by Van Gulick and by Lycan. It is not found in Block's list. However, it should be noted that Block is trying to isolate the different kinds of consciousness that should not be conflated. So, his analysis may be considered as the follow-up step after the differences between conscious and unconscious mental states are acknowledged.

Availability is the second item in our list. Availability is akin to what Block called *access consciousness*, i.e., the content of a conscious state must be available for use by the system. Surprisingly, this item is explicitly stated only by Block. There are the notions of *semantic transparency* and *introspection* in Van Gulick and Lycan, respectively. At a first glance, these may seem to be close substitutes for the concept of availability. However, in both of these notions the emphasis is on the knowledge of the content of the conscious state. Hence this feature may be taken to correspond to the *monitoring consciousness* in Block's distinction.

In addition to the availability of the content of the conscious mental states, conscious beings also have explicit knowledge of this content. The third element in our list acknowledges that there is explicit knowledge of the contents of the conscious mental states. That is, we are aware that we are in a particular state that has a specific content. This can be considered as a kind of higher-order state in the way that Block defines *monitoring consciousness*. Similarly, *introspection* discussed by Lycan, may be seen as such a higher-order process which yields explicit knowledge of the content of the conscious mental states.

We include in our list the concept of qualia as a feature of consciousness. It is clear that we seem to have some subjective raw feels in a conscious experience. So, even if a quale may turn out not to be a real entity, one must nevertheless explain this connection of subjective feels to conscious experiences.

The term *homogeneity* in Lycan and *phenomenal structure* in Van Gulick can be taken as indicating more or less the same property of conscious experiences, namely the unity our conscious experiences seem to display. So, this unity is another feature that should be accounted for by a theory of consciousness. This constitutes the fifth element of our list.

Subjectivity or the Subjective Point of View is the last item on our list. As stated by Lycan and by Van Gulick, conscious experiences have an essential point of view of the subject (termed by these two authors as *subjectivity* and *intrinsic perspectivalness*, respectively).

It is worth noting here that we do not presume consciousness to be an entity with no degrees or gradations. If one thinks that consciousness has degrees or levels, we

Table 1
Features of consciousness.

	Feature or aspect	Remarks
1	Difference	Difference between conscious and unconsciousness mental states
2	Availability	The content of conscious mental states is available to (can be used by) the whole system
3	Explicit and Direct Knowledge	Explicit and direct knowledge of the contents of conscious mental states ("monitoring consciousness")
4	Qualia	Raw feels or sense data that make up the "qualitative features" of an experience
5	Phenomenal Connectedness/Unity	Different modalities of perception are united in a single experience
6	Subjectivity/Subjective Point of View	All conscious experiences are from the viewpoint of the subject; they belong to a single subject

can easily account for that by referring to the fact that at least some of the items in our list, such as Difference and Availability, also permit degrees or levels.

The six features of consciousness that we have identified with help from the philosophy of consciousness literature to date constitute our compiled list. It is shown in Table 1 with the summarized versions of the entries and explanations of what they are.

As will be noticed, we omit some items from the lists given by the three prominent authors in the current literature. Strictly speaking, we do not omit phenomenal consciousness, which is on Block's list. Rather, we follow Van Gulick in dividing it into its constituents. That is, we think that the concept of Qualia, Unity and Subjective Point of View can be taken as the essential components of phenomenal consciousness as identified by Block. Our reasons for our omissions follow.

The first item that does not appear in our list is *self-consciousness*. Subjectivity has surely got something to do with the notion of self. However, self-consciousness, as it is defined by Block, is different from Subjectivity. As it appears in Block's list, self-consciousness requires its possessor to have a concept of self and be able to use this concept. We think that having the concept of self is a different issue from having consciousness, albeit perhaps a closely related one. Clearly, to explain self-consciousness, we need to understand, at a minimum, the concept of self and the concept of consciousness. A theory of consciousness alone is insufficient to fully explain self-consciousness without an accompanying theory of the concept of self.

Secondly, we also omit *creature consciousness* found in Van Gulick's list. Van Gulick states that it may be unrealistic to expect an animal to have certain kinds of mental state, but that we may still want to call this animal conscious. So, according to Van Gulick, proposing that a creature is conscious only if it is capable of having those kinds of mental state may be to set the standard too high. Accordingly, he opposes the idea of explicating creature consciousness in terms of state consciousness, and includes creature consciousness as a separate item in his list (Van Gulick, 1995). However, if a "complete theory of consciousness" is ever developed, then we must but accept the standards of that theory. And creature consciousness should be explicable, we think, by such a theory in terms of a creature's having conscious states. Given all these considerations, there is no need to regard creature consciousness as a separate category independent of state consciousness.

Further, the items *funny facts*, *the ineffability of* "*what it's like*", and the *explanatory gap* that are in Lycan's list are not in our list. We do not think these can be considered as the features of consciousness that must be treated separately. These concepts are almost entirely based on the arguments of Jackson, Nagel and Levine. And those arguments are based on the concept of qualia and the subjectivity of conscious experience. So, it seems plausible to think that if our understanding of the concept of qualia and subjectivity improves, we will also be able to explain *funny facts*, *the ineffability of* "*what it's like*", and the *explanatory gap*. Accordingly, we do not consider it necessary to add these three to our list as separate items, conceptually distinct from the notions of Qualia and Subjectivity.

3. Computer models of consciousness

3.1. What is a model?

We may speak of a one-tenth scale model of a new aircraft design that is being tested in a wind tunnel and a set of computational fluid dynamics equations which is referred to as an aerodynamic model of an aircraft. Both of these models serve the purpose of testing and evaluating the new design. When we speak of climate models, economic models or population dynamics models, there is an implicit notion that these models have something to do with scientific theories. We will not address here such different usages of the term 'model.' However, we must address the different interpretations of the terms 'model' and 'modeling' to the extent that it relates to the concerns of this study.

There are many types of models that can be considered as scientific. Mathematical models and iconic models, which may be physical or computational, clearly have some similarities in the sense that there is a one-to-one correspondence between the elements of the model and the aspects of the phenomena being modeled. A model may be used not only to represent the features of an object, but also the functionality of the object. The architectural model of a bridge is an example of a model representing the features of the actual object. A simulation model, on the other hand, would be an example where the intent is to model some functionality of the object, not just its static elements.

In computer models of consciousness, obviously we are interested in functionality and behavior rather than physical aspects or features. Even if we can identify some elements of consciousness, it is by no means obvious how the human mind works and achieves consciousness. There are various theories about consciousness and how such mechanisms are implemented by the human nervous system, but being a new field, very little has been established as indisputable so far. Accordingly, many computer models of consciousness on the market today are used as investigative tools to assist us in the acquisition of insights that may potentially lead to theories of consciousness, as opposed to mechanically implement well established and generally accepted theories of consciousness.

Before we proceed, let us revisit the model–theory interrelations. Philosophy generally recognizes three types of model–theory interrelations. These are the Received View, the Semantic View, and the Autonomous View (Da Costa & French, 2000). The Received View considers the model to be an often simplified illustration of the theory. This simplification may be valuable in explaining the theory or illustrate its various aspects. Thus the Received View

presupposes that a theory for the phenomena exits, and on this view the theory–model relationship is one-way. The model does not contribute to the theory; it simply illustrates it.

The Semantic View has a slightly different view of what a theory is. It suggests that a theory provides the structure by which our knowledge of a given phenomenon may be organized. The structure provided by the theory is materialized by the models that implement and represent that structure. For instance, a given structure of knowledge may be organized into a set of equations, which one may call a mathematical model describing that theory. Newton's Theory of Gravitation may be summarized, for instance, by his well-known formula. The theory can equally be described verbally or by a computer algorithm that predicts the gravitational forces when given the relevant input. This notion of the relationship between the models and the theory governing them has also been interpreted as the theory just being a set of all possible models that could be realized from the putative structured knowledge.

The last view, which is usually referred to as the Autonomous Model View, is to a great extent a result of recent work based on computational models. In subject areas such as biology and economics, fields in which there is a scarcity of established theoretical models, scientific inquiry attempts to construct computational models to gain insights into the behavior of the underlying complex and often dynamical systems. It is possible to view the autonomous models as preliminary trials from which successful future theories may emerge. In one sense, as we build an autonomous computational model, we are formulating a tentative theory in parallel to it. This approach seems to be successful in the investigation of complex dynamical systems that hitherto have been beyond the reach of formalization by the more conventional approaches.

We see that computational models of consciousness would best be considered as being examples of the Autonomous View. The 1960s saw the beginnings of the flurry of activity in artificial intelligence studies and cognitive sciences. Newell and Simon (1976) proposed the Physical Symbol System Hypothesis. This hypothesis states that having a physical symbol system is necessary and sufficient for general intelligence. The significance of this hypothesis is that it suggests that it is possible to build an intelligent machine. The task at hand is to figure out how exactly the set of symbols and the structure of this machine are to be constructed.

There are basically two schools of thought that manifest themselves as two different main approaches (Cooper & Fox, 2002). The first school, mostly called "traditional symbolic modeling," maintains that artificial intelligence and hence cognition can be achieved by a set of rules which operate on a set of symbols. The second school of thought is based on artificial neural networks, sometimes referred to as "connectionist models." Artificial neural networks are inspired by the structure of the nervous system, where there are many interconnections among the neurons conducting electrical signals which exhibit complex patterns. Supporters of the school point out that since the brain seems to be built as an inter-connected set of neurons, it is justifiable to construct theories of cognition based on the same structure.

Both schools of thought have their strengths and shortcomings. While symbolic rule-based systems are easily written and comprehended, it becomes rather difficult to implement learning with this approach. Neural networks, on the other hand, are capable of incorporating learning. However, since their input–output relations depend on a set of weights distributed over a large number of neurons, the exact functioning mechanism becomes somewhat obscure to casual inspection. The neural network is thus a more holistic instrument while the symbolic rule-based approach tends to be closer to the more customary formal systems. Besides the pure symbolic approach and the connectionist approach, there are also hybrid models that attempt to combine the better aspects of the two.

Although the desire to build thinking machines gave impetus to the computer models of cognition, there are differences in terms of the purpose and motivation among the various studies. While some studies set as a goal the construction of an artificially intelligent machine, others aim to model intelligence to understand the underlying mechanisms of human cognition.

Another way of classifying computer models of cognition stems from whether the models are to be used as practical end products or as academic instruments for scientific inquiry. There exist practical implementations such as IDA and LIDA that take on useful responsibilities. Scientifically oriented models such as Clarion, on the other hand, are mostly used as investigative tools to enhance our understanding of cognition. Of course, the practical models could be useful in scientific inquiry as well, and similarly, versions of the scientific models could form the basis of future practical products.

In cognitive science one considers different domains of cognitive faculties. For example, faculties such as vision, memory, decision making may be considered as different domains. Another classification of computer models can be achieved along the lines of domain specificity. Some models propose architectures or frameworks that could potentially be customized to deal with any domain. Others are domain- or task-specific. For instance, they may be specific to the domain of attention. Newell (1990) makes an argument that if we are to understand human cognition we must focus on architectural models that are domain-independent, since these are more capable of conveying information about the general notion of cognition. On the other hand, if task-specific models were to converge in structure and behavior, they would have significant implication as the structure of cognition.

3.2. Computer models of consciousness

This section reviews some of the better known the self-claimed computer models of consciousness that have

appeared in the literature. (We explain what is meant here by "self-claimed" in footnote 1.) We restrict our review to those models that have been implemented and their behavior examined.

3.2.1. Clarion

Clarion (for "Connectionist Learning with Adaptive Rule Induction ON-Line") (Sun, 2003, 2006) is a hybrid architecture that involves both connectionist and rule-based elements. Clarion recognizes implicit and explicit knowledge. It uses distributed networks to represent sub-symbolic implicit knowledge and localist networks are employed in a symbolic way to represent explicit knowledge. This leads to a two-level architecture, where the top level, or the explicit knowledge level, and the bottom level, or the implicit knowledge level, are modeled by different modules. Being a flexible and extensible framework, Clarion employs several subsystems such as the Action-Centered, Non-Action-Centered, Motivational and Meta-Cognitive Subsystems. Each subsystem contains top level and bottom level modules to handle explicit and implicit knowledge, respectively.

Clarion implements Explicit Knowledge by using rules and chunks (dimension/value pairs), tracks the "state of the world," implements a working memory,[5] and goals, all of which use dimension/value pairs. Rules use chunks and scan the dimension/value pairs present in the state of the world, working memory and goals. If all of the dimension/value pairs in the condition part of a rule are present, then the rule is triggered. As a result, the dimension/value pairs in the action part of the rule are introduced to the working memory. At the top level, the output dimension/value pairs are generated from the input dimension/value pairs through this rule mechanism. The bottom level also generates output dimension/value pairs from input dimension/value pairs. However, the mechanism here is more implicit. A back-propagation network rather than rules is used. Since both the top level and bottom level knowledge is represented by dimension/value pairs, these two levels can freely exchange the represented knowledge. The bottom level uses Q-learning. There are many alternate learning mechanisms at the top level. Some of these are specific to a given subsystem. The learning mechanisms include top-down assimilation, imitative learning, rule extraction, and independent rule learning.

[5] The term 'working memory' was used for the first time in the work of Baddeley and Hitch dated 1974 (as cited in Baddeley, 2003). In psychology, "the concept of working memory proposes that a dedicated system maintains and stores information in the short term, and this system underlies human thought processes" (Baddeley, 2003, p. 829). The core idea of working memory has been preserved from that time, although there have been some modifications to the original model. One relevant such modification is the introduction of an "episodic buffer" to the system. Through this modification, working memory now becomes also the seat of consciousness (Baddeley, 2003, p. 836). For a review of the studies on working memory see Baddeley (2003).

Considering the distinction between conscious and unconscious processes included in our list, Clarion, with its two-level architecture, uses the lower level neural networks to model the unconscious representations and processes, while the upper rule-based level corresponds to the conscious representations and processes. In this respect, Clarion models the two and clearly delineates them.

The next item on our list, Availability, does not hold in Clarion. Sun (1999) explicitly refers to *access consciousness* and defines it as the "direct availability of the mental content for access" (p. 534). He also proposes that the difference in the kind of representation of conscious content, i.e. the explicit representation, is enough to account for the direct availability. Although information can be shared among the subsystems, no attempt is made to elevate the top-level conscious processes more than any other information. Lower-level information is shared as much as upper-level information, albeit indirectly. However, the nature of the indirectness is not quite clear since all subsystems have a two-level structure, where the processes and representations at the bottom level can affect the input and output processes without any interaction with the top level. Thus, we must conclude that Availability is not completely addressed by Clarion.

As stated by Sun (1999), the conscious representations are modeled by localist networks to capture their explicit nature. Also, there are some meta-level processes serving as reasoning mechanisms that utilize the explicit representations at the top level. So, Clarion fulfills the third item in our list.

The last three items on our list are Qualia, Connectedness of experience, and Subjectivity, all of which are related to phenomenal consciousness. As explained earlier, phenomenal consciousness is interpreted in different ways and discussed with its different facets in the literature. This is why we explicitly list Qualia, Connectedness of experience, and Subjectivity as separate items, although other researchers have used the term 'phenomenal consciousness' to roughly refer to any one of these.

Clarion claims to address phenomenal consciousness. More specifically, Clarion states that since there is a representational difference between the upper and lower level processes, the model captures phenomenal consciousness. Clearly, such division is not sufficient to address Connectedness of experience or Subjectivity. What Clarion can at best claim to address may be the item Qualia. But, although the separate representation of upper and lower level processes may be a plausible initial step to explain qualia, simply implementing such a division is not, in our view, sufficient to model qualia.

3.2.2. LIDA (Learning IDA)

IDA (for "Intelligent Distribution Agent") (Franklin, 2000, 2003) is somewhat different from the other models in that it has been given a specific practical task to perform: IDA is used by the US Navy to assign specialists to new posts. IDA must communicate with the sailors through

e-mail using natural language. It must access the databases to see what is needed and what is available. It must also observe the rules and regulations of the Navy in making these decisions. However practical the end task is, IDA is nonetheless claimed to be an architecture that models a "conscious" mind. The term 'consciousness' is put in quotations here, following the developers, since there is no claim to model all aspects of consciousness.

IDA implements the global workspace theory as an autonomous software agent. An autonomous agent is defined as "a system situated in, and part of, an environment, which senses that environment, and acts on it, over time, in pursuit of its own agenda" (Franklin & Graesser, 2001). Autonomous software agents are claimed to facilitate cognitive sciences by promoting several hypotheses. Working with such agents, therefore, provides insights into the workings of the mind.

The global workspace theory put forth by Baars (1988) proposes that consciousness involves several processes, some of which enter a global workspace. Once in the global workspace its content becomes available to other elements and processes. The processes in the global workspace decay over time, which allows the global workspace to be a dynamical mechanism. Viewing the many processes as separate agents, one may also consider the model to be a hierarchical multi-agent system. The developers of IDA view a software agent that implements the global workspace theory to be a "conscious" software agent.

IDA is a hybrid system that combines both symbolic and connectionist elements. The components responsible for the so called high-level abstractions, such as behavior and emotions, are combined with low-level elements, known as codelets. Each codelet is a piece of code that performs a specific task. Codelets perform these tasks independently and concurrently. Codelets constitute the multiple agents of IDA. Those codelets that successfully identify some aspect of the input activate nodes of a slipnet. When the slipnet settles, the output corresponds to the derived meaning of the inputs.

IDA employs what is called a sparse distributed memory (SDM). SDM is content-addressable, which is claimed to be well-suited for long-term associative memory. Retrieval from SDM is implemented as an iterative process. If the target is not found within a predetermined interval, IDA generates the response "I don't know."

Some codelets, in turn, activate nodes in the global workspace. There are several components of this architecture: there is a coalition manager, a spotlight controller, a broadcast manager, and a set of attention codelets. Codelets continuously scan the inputs to see if there is anything new. The signals from the codelets that identify a new input are combined by an attention codelet. The attention codelet and the codelets it is associated with comprise a coalition. The strength of the coalition depends on how well the codelets match the input to the conditions they are to identify. The coalition manager regulates this process. Each coalition then competes for consciousness. The spotlight controller determines which of the coalitions are to be elevated to the global workspace. Finally, the coalition elevated to the global workspace is broadcast system-wide by the broadcast controller.

Action selection in IDA is implemented by behavior codelets. A behavior is similar to an (if-then type) production rule that has conditions and associated actions. Behavior codelets in IDA also take into consideration the strength of the conditions in determining the actions.

A more recent version LIDA, i.e. Learning IDA (Baars & Franklin, 2009; Snaider, McCall, & Franklin, 2012), has similar working mechanisms,[6] at least in so far as those mechanisms that are critical to our assessment are concerned. However, there is a new subcomponent of the workspace in LIDA, called "Conscious Contents Queue," that is of particular importance. Only the contents that are broadcast are added to this component, where the most recent contents are represented as the first elements. Moreover, the codelets in the workspace have a direct access to this subcomponent.

The Difference between conscious and unconscious mental states in our list is addressed by the global workspace of LIDA.[7] In accordance with Baars' theory, information that is elevated to the global workspace can be considered to represent the content of conscious mental states. In this respect, LIDA fulfills the first item on our list.

The next item, Availability, also holds in LIDA. Recall that some information is broadcast to the entire system. This makes it available to any component within the model. Thus, LIDA fulfills the second item on our list as well.

The implementation of the subcomponent Conscious Contents Queue is a promising step towards the fulfillment of our third element, since the contents of recently conscious states are explicitly represented in this subcomponent. Yet, since this subcomponent is currently utilized only for the representation of time, the mechanisms for it to be used in higher level cognitive capacities (e.g. for monitoring) are not specified.

The last three items in our list are Qualia, Connectedness of experience, and Subjectivity. Being task-oriented, LIDA makes no claim or attempt to model any one of these. Curiously though, at one point, the statement is made that the broadcast "is hypothesized to correspond to phenomenal consciousness" (Franklin, 2000). Beyond providing a good example of how in the literature phenomenal consciousness and Availability are sometimes confused, we find little value or credence in this claim. In all fairness, however, LIDA may have the beginnings of some

[6] Yet, as the change in the name suggests, there are some differences between the two with respect to the implementation of mechanisms for learning.

[7] Since the mechanisms of IDA and LIDA are similar and LIDA is the more recent version, we will be referring to LIDA in the remainder of this article.

aspects of what we called connectedness of experience. Specifically, we mentioned the unity of conscious experience. The coalition manager and the related structures of LIDA may be used to implement such aspects into a computer model of consciousness.

3.2.3. ACT-R

ACT-R (short for "Adaptive Control of Thought – Rational") (Anderson, 1996; Anderson et al., 2004), and its predecessor ACT ("adaptive control of thought") are open architectures that can be programmed to perform several tasks. ACT-R is somewhat different from other computer models of consciousness. At the outset, ACT-R does not make an explicit attempt to model consciousness per se. It proposes an architecture for the mind without the stated objective to delineate consciousness. It is interesting, however, that the resultant model displays elements of consciousness present in other models which explicitly set out to address consciousness (Taatgen, 2009).

ACT-R attempts to integrate several modules with a core production system. A set of modules are implemented and made available. Users may insert their own production rules and experiment with the ACT-R architecture. Critical to the concept of a module is that each module contains a buffer. Just as a person may see many objects in her field of vision but focuses her attention on a specific object, the buffers selectively hold the information most relevant to the task at hand. ACT-R allows these buffers to interact, with the involvement of the production rules. Specifically, the production rules scan the buffers, and the rules that fire produce results that are also kept in the buffers. The buffer contents thus dynamically change as new inputs are received and the production rules are applied.

Many of the processes are carried out in parallel. However, some processes must be performed in series. The buffer content is limited to a single declarative unit of knowledge, which is referred to as a "chunk" in ACT-R. Accordingly, only a single memory can be deposited or retrieved at a time. Similarly, production rules are fired one at a time, at each cycle. The integration comes from the fact that each module also deposits information about its activities into its buffer. The procedural memory module that contains the production rules has access to this information. The production system can detect patterns of module behavior and fire rules depending on the observed patterns. This provides a higher-order coupling among the modules, which gives rise to the integrated nature of ACT-R.

The distinction between the information in the modules and in the buffers fulfills the requirement of our first item, namely the distinction of conscious and unconscious mental states.

ACT-R also scores high on our second item: the Availability of conscious information to other processes. The production system of ACT-R has access to all of the information in the buffers. Thus, once placed in a buffer, the information becomes available to all processes.

The third item, Explicit and Direct Knowledge does not correspond to any part of ACT-R. We thus conclude that ACT-R does not fulfill this item on our list. The remaining three, viz. Qualia, Connectedness of experience, and Subjectivity all relate to phenomenal consciousness. Again, ACT-R does not address these elements at all.

3.2.4. Neuronal Work Space Model (NWS)

Dehaene and Naccache (2001) propose a neuronal work space theory of consciousness, basing it on three empirical observations: the existence of unconscious cognitive processing, the necessity of attention in conscious processing, and the requirement of consciousness in some particular effortful tasks. It is claimed that these empirical observations can be explained by a theory that takes as a launching platform Baars' global work space theory,[8] and by imposing the constraints that follow from the structure of the brain.

Accordingly, they postulate three main theoretical considerations. Firstly, they propose that there are some processes that occur in specific brain areas, and can operate without an attentional mechanism or availability to the whole system. On the other hand, as their second theoretical claim, they propose a distributed neuronal "workspace" network that connects several of these brain areas. Also, the level and duration of activation are important for a process to gain access to this neuronal workspace. The top-down amplification of conscious states via attentional processes is the third theoretical postulate.

There are several implemented, yet simplified, neural network models of this theory that simulates relevant human behavior. One such network, reported in Dehaene, Sergent, and Changeux (2003), models the phenomenon known as attentional blink.[9] In the model, there are two sub-networks, one for each task, and the two sub-networks are linked by inhibitor connections, each suppressing the activation of the nodes of the other. The sub-networks have also re-entrant connections that enable the stability and sustain the activation in each sub-network for a while. Whichever sub-network receives the input first starts its computation ahead of the other and becomes the first to activate the inhibitors. The second input is processed in the usual way until it reaches the stage where its further progress is inhibited by signals from the earlier activated sub-network. In fact, when two inputs in rapid succession are received, only the first reaches the termination, while the second is suppressed. This model successfully mimics the phenomenon of attentional blink.

In this model, the distinction between conscious and unconscious representations is realized both by amplified activation and access to neuronal workspace. In the simulation, these two constraints are realized by the

[8] See Section 3.2.2 for a brief explanation of Baars' theory.

[9] Attentional blink refers to the situation where two inputs are issued at rapid succession. While the subject focuses her attention on the first task, she misses the second input. That is to say, her attention blinks during the second input and causes her to miss the second input entirely.

re-entrant connections and inhibitory connections in higher levels, respectively.

Access to neuronal workspace accounts also for the Availability item in our list. Once the activation of a particular process is elevated to the workspace, it can influence the activities of other processes.

The third item, Explicit and Direct Knowledge of the contents of the conscious states, is not considered by this model. Nor do the authors claim to have implemented such a feature.

The remaining three items on our list were Qualia, Connectedness of experience, and Subjectivity. These are not addressed separately by the model. Yet, Dehaene and Naccache (2001, p. 30) remark on the difference between access consciousness and phenomenal consciousness, as these concepts are formulated by Block. They state that this difference may correspond to the difference between the processes that are activated but cannot gain access to global workspace since they lack attentional amplification, and those that make it into the workspace. According to Dehaene and Naccache, the former processes, in a sense, are potentially conscious. Note however, that, if a process cannot gain access to the workspace, then it is not available to the whole system. But such capability is essential for access consciousness. Thus, it would be rather unjustified to regard the model as fulfilling the requirements of access consciousness.

3.2.5. ART

ART ("Adaptive Resonance Theory") (Grossberg, 1987, 2007) is a framework for developing neural networks. It models human cognitive processes. The basic motivation for this framework is to propose a learning procedure that is different from, and more biologically plausible than, back-propagation. It is also demanded from the framework that it can solve at least the problem known as the "stability-plasticity dilemma."[10] Fundamentally, ART is a matching process that is mediated by an attentional-orienting system.

In an ART network, lower layers represent the features, whereas the higher layers represent the groups, i.e. chunks, of these features. When an input pattern activates a particular layer in the network, the bottom-up feed-forward connections create a pattern of activation that is determined by the weights of the connections at a higher level. Then, through the top-down feedback connections, another activation pattern is generated. If this last pattern and the first pattern are matched, that is, if their difference is smaller than a predetermined "vigilance parameter," the activity of the ongoing bottom-up and top-down activations results in a "resonance state." The resonance states correspond to the states where a particular activation pattern becomes stable. The subsystem that is responsible for this matching procedure is the attentional subsystem. On the other hand, it the match does not occur, that is, if the difference between the initially activated pattern and the pattern activated by the top-down connections is bigger than the vigilance parameter, then the activity at the higher level is inhibited through the orienting subsystem. This inhibition leads to a search for a better matching activation pattern at the higher level. If no such pattern exits, then a new one is created. The duration of activation is long enough to affect the weights of the connections between two layers only in resonance states. So, learning occurs only when there are resonance states.

Given the above considerations, the framework proposes a particular theory that binds consciousness, learning, expectation, attention, resonance, and synchrony. The name, CLEARS, is an acronym constructed from these six ingredients (Grossberg, 2007). Thus, according to this theory, learning takes place through the resonance states that are a result of the matching process of new experiences with previously learned expectations, all mediated by attention. Also, these long enduring resonance states are hypothesized to be conscious states, which are realized by the synchronized oscillations in the brain.

With respect to the first item in our list, ART proposes that the difference between conscious and unconscious states is the conscious states' being attention oriented resonance states. Through the matching procedure some intermediate activation patterns also occur. Yet, since only the matches that are strong enough are capable of leading to a resonance state, these intermediate states remain unconscious.

Availability is not directly addressed in ART. The existence of top-down and bottom-up connections between all pairs of layers, in a sense, implies that all activation patterns are affected and can affect the rest of the network. However, the prolonged duration of conscious states, i.e. the resonance states, may be a starting point for the implementation of a mechanism that can explain the distinct availability of conscious states.

There is no proposed mechanism in ART which can maintain the internal representations of the network. So, the model does not fulfill our third item, namely the implementation of Explicit and Direct Knowledge.

The only item that is partially addressed in ART regarding the remaining three items of our list, viz. Qualia, Phenomenal Connectedness of experience, and Subjectivity, is the connectedness of the experience which is related to the unity of the conscious experience. Grossberg (2007, p. 1047) proposes that the resonance states bind the distributed features into more coherent higher level representations. This binding of features may be enough to form a unity in certain specific modalities. However, it is not enough to explain the connectedness and the unity of various different modalities in a conscious experience.

3.2.6. GMU-BICA

GMU-BICA ("George Mason University – Biologically Inspired Cognitive Architecture") (Samsonovich & De Jong, 2002; Samsonovich, De Jong, & Kitsantas, 2009) is a hybrid cognitive architecture that is developed to model especially higher level human cognitive abilities. It is a

[10] As it is stated by Grossberg (1987, p. 30) the problem is "How can a learning system be designed to remain plastic in response to new events, yet also remain stable in response to irrelevant events?".

recently developed architecture and it uniquely focuses on "self." Yet, as it is explicitly stated by the modelers, the notion of self in this architecture is different from the philosophical one, and is "a structureless, abstract token to which contents of mental states can be attributed, rather than the cognitive system itself or any of its observable aspects" (Samsonovich et al., 2009, p. 114). One important theoretical proposal of GMU-BICA is the definition of a mental state. A mental state is taken to consist of not only the content, but also a subjective perspective.

The contents are realized by "schemes" in the architecture. Each scheme has a predefined number of attributes that can assume specific values according to past or present experiences of the agent. The generic schemes are stored in the "Semantic Memory" component. Currently instantiated schemes, i.e. schemes that have specific values for their attributes, are said to form the "Working Memory" component. Each mental state has a limited number of schemes, along with a label that points to the subject of the mental state. The interactions occur only among the schemes that are the contents of the same mental state. Separate mental states can only copy a scheme (with appropriate modifications of its attributes).

The only mental state that is required to always be present in the "Working Memory" is the "I-Now." This state has the perspective of the agent and contains schemes that have the values that represent the current states of affairs. "I-Now" has also privileged access to the "Input/Output" component. Although "I-Now" is the only required mental state, there may also be other mental states representing, for example, past experiences of the agent (labeled as "I-Previous"), or current experiences of other agents (labeled as "He-Now"). There is also one mental state of particular importance that may be present in the "Working Memory," namely "I-Meta." This mental state can modify the contents of mental states that have the perspective of the agent's self, i.e. those states labeled as "I-".

What GMU-BICA specifically proposes as the difference between conscious and unconscious mental states is not so clear. Samsonovich et al. (2009, p. 115) state that "it is a scholastic question whether only the content of I-Now – or the entire content of the working memory should be associated with 'consciousness' of the agent." Yet this so called scholastic question cannot be easily ignored if one aims to model human cognitive faculties.[11] Actually it seems more convenient to attribute consciousness to "I-Now" mental states, since mental states like "He-Now" may also be present in the "Working Memory." These mental states cannot be taken as the agent's mental states that represent the mental states of other agents because they have the subjective perspective of another agent (as it is indicated by their label). In light of these considerations, the architecture only partially fulfills the first item in our list.

As stated above, the "I-Now" states have also a privileged access to the "Input/Output" component. Also, the fact that "I-Meta" can modify the contents of the "I-Now" implies that the contents of these states are in a sense available to the other states. So, the architecture proposes a mechanism for the Availability item in our list.

The only candidate that might explain the third item in our list, viz. Explicit and Direct Knowledge, is the "I-Meta." However, the above considerations show that the peculiar property of this state is its capability to modify some other mental states. This, we think, offers a mechanism for the list item Availability, but not for the item Explicit and Direct Knowledge of the conscious content.

Among the last three items, the only item that is addressed by GMU-BICA is Subjectivity. This is not surprising, since the initial motivation of the architecture is to delineate the notion of self. However, merely labeling mental states as belonging to certain perspectives of subjects is not enough without specifications of the particular mechanism. Also, taking the labeling as the sole explanation seems rather implausible due to the fact that there are other mental states labeled as the perspectives of other agents in the "Working Memory."

4. Conclusion

We evaluated six implemented computational models of consciousness according to the six features of consciousness identified by philosophy. The following table summarizes the results. In the table, the features a model successfully implements are denoted by a plus (+). The cells that correspond to the features not addressed by the respective model are left blank. Some of the cells contain question marks. These correspond to partial models or efforts that may need further amplification.

When Table 2 is viewed column-wise, it reveals information about how successful the corresponding model is in satisfying the features in our list. Here, we see that LIDA fares better. When viewed row-wise, Table 2 shows how well the currently implemented computational models address each feature in our list. On that score, Difference and Availability seem to be addressed more than the other features. By contrast, features such as Phenomenal Connectedness and Subjectivity are hardly implemented by these models. Recall that the less represented features are related to phenomenal consciousness.

[11] It is true that some questions can be ignored if one aims to develop a cognitive architecture that can mimic the human behavior just for practical purposes and does not try to construct any theoretical framework that purports to explain how these behaviors are realized in humans. Indeed, the initial motivation of GMU-BICA seems to be along this line (see Samsonovich & De Jong, 2002). Although it is "biologically inspired" right from the beginning, an inspiration does not necessarily constitute criteria for justification. Yet, once it is claimed that an architecture is consistent with psychological data and is capable of making predictions (see Samsonovich et al., 2009), one has a right to expect not only correspondence with behavioral data, but consistency of the concepts employed by the model.

Table 2
Results of the study.

	Clarion	LIDA	ACT-R	NWS	ART	GMU-BICA
Difference	+	+	+	+	+	?
Availability		+	+		?	+
Explicit and Direct Knowledge	+	?		+		
Qualia	?	?		?		
Phenomenal Connectedness/Unity		?			?	
Subjectivity/Subjective Point of View						?

The presence of many question marks in Table 2 also conveys an important message. Although the models attempt to address these features, there is much room for improvement before we can accept that the models completely and successfully address and implement these aspects. Our observation earlier that LIDA implements more features than the others takes into account also the question marks which we optimistically include in the final score of the features it implements.

Moreover, the abundance of the question marks and empty cells signifies, in our view, opportunities for future work towards closing the gap between computer models and philosophical insights. In fact, which models will address the most features, and hence become the most complete in the future, depends mostly on how many of the aspects labeled with a question mark they improve in implementing. In other words, this study brings to the discussion table, the value of reviewing computational models not only to evaluate and rank them, but also to suggest what further aspects the new computational models may wish to implement.

In light of these results, a few points become evident. First, we would like to submit that modeling consciousness has a synergistic effect on consciousness studies. We believe that as more models are developed, more insights are obtained into the workings of the mind. However, there are a precious few models reported as yet in the literature. It is plausible to suppose that as the number of computer models increases by one or two orders of magnitude, our insights and understanding will also be enhanced significantly. Some of the computer models provide hints for the initial construction of possible theories of the related elements of consciousness. Further modeling efforts can be expected to refine and hone the theories towards successful explanations of the workings of the mind. In this sense, we regard the present computer models of consciousness as following the Autonomous View of modeling, where the theory and the model are somewhat independent.

Another observation to be made is that the first three items on our list (namely, the Difference between conscious and unconscious mental states, Availability, and Explicit and Direct Knowledge) are all taken into consideration by at least one of the computer models. More work in these areas should be useful in further developing our understanding. By contrast, the last three items are not addressed by the current models as prevalently as the first three. These last three items all relate to phenomenal consciousness.[12] In fact, it seems that there is quite a bit of confusion concerning phenomenal consciousness. One may even be bold enough to view phenomenal consciousness as the "catchall" category, into which other unexplained phenomena are deposited. Especially as it concerns aspects of phenomenal consciousness, there seems to be a genuine need for philosophy, computer science, and other disciplines to cooperate. Such cooperation should aim to carefully identify the components and elements of phenomenal consciousness so that different disciplines agree on the concept of these terms. The identification of the confusion surrounding phenomenal consciousness was a significant result of this study.

We would like to make a pragmatic suggestion for future modeling efforts. As can be seen from the table above, the aspect of Subjectivity, although a most important feature identified by philosophy, is partially addressed in only one of the current models in the literature. New models would be well advised to make attempts to address Subjectivity, for without this feature, a complete and comprehensive understanding of consciousness seems unattainable.

As a final note, we want to re-emphasize the point made earlier in the introduction of this paper. The evaluation of computer models in this study is based on the literature and on the claims of their respective developers, rather than on hands-on modeling experience. We see this as a particular weakness of the study. Yet, this weakness also suggests a path to improvement. One may, for example, try to model the particular features of consciousness in the environments of particular computer models, as a natural extension of this study.

Acknowledgments

We would like to thank Sencer Yeralan for his participation in the preliminary discussions that led to this work, and for his comments to the earlier drafts of this paper. We would also like to thank the three anonymous referees

[12] In fact, it may be said that it is better to pack the three items together in a single item, since most of the question marks appear in the corresponding cells. However, we think otherwise. The question marks indeed indicate that we need a more careful analysis of these three items. But uniting them introduces the potential danger of focusing on only one of them and ignoring the others. In fact, this is exactly the case in some models, which claim to capture phenomenal consciousness.

for their valuable comments to the first version of this paper. Their constructive comments led to significant improvements to the paper.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*, 355–365.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060.

Armstrong, D. (1997). What is consciousness? In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 712–728). Cambridge: MIT Press.

Baars, B. (1986). *The cognitive revolution in psychology*. New York: The Guilford Press.

Baars, B. (1988). *A cognitive theory of consciousness*. <http://vesicle.nsi.edu/users/baars/BaarsConsciousnessBook1988/>.

Baars, B., & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness, 1*(1), 23–32.

Baddeley, A. D. (2003). Looking back and looking forward. *Nature Reviews: Neuroscience, 4*, 829–839.

Block, N. (1997). On a confusion about a function of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 375–416). Cambridge: MIT Press.

Cooper, R., & Fox, J. (2002). Modelling cognition. In R. Cooper (Ed.), *Modeling high-level cognitive processes*. Mahwah: Lawrence Erlbaum Assoc..

Da Costa, N., & French, S. (2000). Models, theories, and structures: Thirty years on. *Philosophy of Science, 67*, 116–127.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *PNAS, 100*, 8520–8525.

Descartes, R. (1973). *The principles of philosophy (Translated by E. Haldane and G. Ross)*. Cambridge: Cambridge University Press (Original work published 1644).

Franklin, S. (2000). A "Consciousness" based architecture for a functioning mind. In A. Sloman (Ed.), *Proceedings of the symposium on designing a functioning mind*. <http://www2.dcs.hull.ac.uk/NEAT/dnd/visions/Proposals/stan.pdf>.

Franklin, S. (2003). IDA: A conscious artifact? *Journal of Consciousness Studies, 10*, 47–66.

Franklin, S., & Graesser, A. (2001). Modeling cognition with software agents. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the cognitive science society*. Mahwah: Lawrence Erlbaum Assoc..

Grossberg, S. (1987). Competitive learning: From interactive attention to adaptive resonance. *Cognitive Science, 11*, 23–63.

Grossberg, S. (2007). Consciousness CLEARS the mind. *Neural Networks, 20*(9), 1040–1053.

Güzeldere, G. (1997). The many faces of consciousness: A field guide. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 1–67). Cambridge: MIT Press.

Hameroff, S. R., & Penrose, R. (1995). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation, 40*(3), 453–480.

Jackson, F. (1997). What mary didn't know. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 567–571). Cambridge: MIT Press.

Leibniz, G. W. (1989). *Philosophical essays. Edited and translated by R. Ariew and D. Graber*. Indianapolis: Hackett Publishing Co. (Original work published 1765).

Levine, J. (1997). On leaving out what it's like. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 543–556). Cambridge: MIT Press.

Lycan, W. G. (1997). Consciousness as internal monitoring. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 755–772). Cambridge: MIT Press.

Lycan, W. G. (1999). Plurality of consciousness. <http://www.unc.edu/~ujanel/CogThs.html>.

McGovern, K., & Baars, B. J. (2007). Cognitive theories of consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness*. Cambridge: Cambridge University Press.

Nagel, T. (1997). What is it like to be a bat? In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 519–528). Cambridge: MIT Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge: Harvard University Press.

Newell, A., & Simon, H. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Rosenthal, D. M. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 729–754). Cambridge: MIT Press.

Samsonovich, A. V. & De Jong, K. A. (2002). Designing a self-aware neuromorphic hybrid. <http://www.aaai.org/Papers/Workshops/2005/WS-05-08/WS05-08-011.pdf>.

Samsonovich, A. V., De Jong, K. A., & Kitsantas, A. (2009). The mental state formalism of GMU-BICA. *International Journal of Machine Consciousness, 1*(1), 111–130.

Snaider, J., McCall, R., & Franklin, S. (2012). Time production and representation in a conceptual and computational cognitive model. *Cognitive Systems Research, 13*(1), 59–71.

Sun, R. (1999). Computational models of consciousness: An evaluation. *Journal of Intelligent Systems, 9*, 507–562.

Sun, R. (2003). A tutorial on CLARION 5.0. <http://www.cogsci.rpi.edu/~rsun/sun.tutorial.pdf>.

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In R. Sun (Ed.), *Cognition and multi-agent interaction*. Cambridge: Cambridge University Press.

Taatgen, N. A. (2009). Consciousness in ACT-R. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness*. Oxford: Oxford University Press.

Tonini, G. (2004). An information integration theory of consciousness. *BMC Neurosicence, 5*(42), 3.

Tye, M. (2000). *Consciousness, color, and content*. Cambridge: MIT Press.

Van Gulick, R. (1995). What would count as explaining consciousness? In T. Metzinger (Ed.), *Conscious experience*. Paderborn: Ferdinand Schöningh.