MODELING CONSCIOUSNESS:
A COMPARISON OF COMPUTATIONAL MODELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

SELVİ ELİF GÖK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

SEPTEMBER 2009

Approval of the Graduate School of Informatics

_____

Prof. Dr. Nazife Baykal

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Prof. Dr. Deniz Zeyrek

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. Erdinç Sayan

Supervisor

Examining Committee Members

Assist. Prof. Dr. Bilge Say (METU, COGS) _____

Assoc. Prof. Dr. Erdinç Sayan (METU, PHIL) _____

Assist. Prof. Dr. Annette Hohenberger (METU, COGS) _____

Assoc. Prof. Dr. Ayhan Sol (METU, PHIL) _____

Dr. Ceyhan Temürcü (METU, COGS) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this wok.

Name, Last name : **Selvi Elif Gök**

Signature : _____

ABSTRACT

MODELING CONSCIOUSNESS:
A COMPARISON OF COMPUTATIONAL MODELS

Gök, Selvi Elif
M.S., Department of Cognitive Science
Supervisor: Assoc. Prof. Dr. Erdinç Sayan

September 2009, 49 pages

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 detailed and critical review of the literature of philosophy of consciousness, this study constructs a composite and eclectic 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 computational models of consciousness, specifically CLARION, IDA,

ACT-R and model proposed in the Cleeremans' review and study. The computational models studied are evaluated with respect to each identified aspect and feature of consciousness.

Keywords: Philosophy of consciousness, Computational cognitive modeling, Clarion, IDA, ACT-R

ÖZ

BİLİNCİ MODELLEMEK:
BİLGİSAYAR MODELLERİNİN BİR KARŞILAŞTIRMASI

Gök, Selvi Elif
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yoneticisi: Doç. Dr. Erdinç Sayan

Eylül 2009, 49 sayfa

Günümüzde bilinç calışmalarında yoğun bir artış gözlenmektedir. Şu ana kadar elimizde bilincin işlevsel bir tanımı olma da, felsefe alanında bilincin çeşitli yönleriyle ilgili olduğu düşünülen bir takım özelikler ve nitelikler belirlenmiştir. Diğer yandan, bilincin bilgisayar modellerinin geliştirilmesi konusunda da bir takım yeni çalışmalar vardır. Bu modellerin, bilincin bir ya da daha fazla özelligini başarıyla modellediği öne sürülmektedir. Bu çalışma varolan bilinç modellerinin değerlendirilmesi için, bu modellerin bilincin felsefece belirlenmiş özellik ve nitelikleri modellemekte ne kadar başarılı olduğuna bakamaktadır. İlk olarak bilinç felsefesinin detaylı bir incelemesinden sonra, bilincin özeliklerinin birleşik ve eklektik bir listesi geliştirilmiştir. Daha sonra bu liste ışığında, bilinci modellediği geliştiricileri tarafından öne sürülen bilgisayar modelleri incelenmiştir. Bu modeller, CLARION, IDA, ACT-R, ve Cleeremans'ın önerdiği bilinç modelleridir. İncelenilen modeller oluşturulan listedeki herbir özelliği ne kadar açıkladıkları bakımından değerlendirilmişlerdir.

Anahtar Kelimeler: Bilinç felsefesi, Bilgisayarlı bilişsel modelleme, CLARION, IDA, ACT-R

ACKNOWLEDGMENTS

I am deeply indebted to my thesis advisor Dr. Erdinç Sayan, not only for his direction regarding this work, but also for his guidance, kindness, and patiences throughout my undergraduate and graduate studies. I thank Dr. Sencer Yeralan for proof-reading the thesis, many times, and suggesting ideas that helped me to clarify the various points. I thank the members of my thesis supervisory committee, Dr. Bilge Say, Dr. Annette Hohenberger, Dr. Ayhan Sol, and Dr. Ceyhan Temürcü for their constructive comments, which greatly improved the quality of this work.

I am grateful to my family who always make me feel unconditionally loved and supported.

Finally I would like to acknowledge TUBITAK, whose National Scholarship Program for M.Sc. Students, Program 2210, provided the fellowship during my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

CHAPTER 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. Although there is no mature framework at the present for studying consciousness, philosophy has come to identify a preliminary 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 seem to be many alternative views of consciousness. This study takes a look at the 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. In this respect, this study combines philosophy and computer science in reviewing recent work in consciousness research in both fields.

As a plausible substitute to evaluating how well the current computer programs model consciousness, this study examines how they fare in modeling those

1 The reason we call them 'self-claimed' is to indicate the point that recent literature contains works from developers of computer models, who claim that these models can capture certain aspects of consciousness. See the relevant sections of the models in Chapter 4 for details.

aspects and features of consciousness identified by philosophy. 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.

Although there is no work in the literature that duplicates our study, there have been a few reported investigations and evaluations of computational models of consciousness from the viewpoint of a list that at least contains philosophical aspects. Most notably, Coward and Sun (2004), and Sun and Franklin (2007) based their evaluations on the list of features submitted by Block (2007b). These authors are also the developers of the computer models Clarion and IDA. Their evaluations are somewhat arguably and understandably closer to the viewpoint of their own models. In this respect, these studies do not completely provide a comprehensive survey into the state of the art of computer models of consciousness. Seth (2009; also see Clowes and Seth, 2008) starts with a proposed list of features. He then gives a survey of simulation models of consciousness, each of which corresponds to successfully implementing a feature in his list. Note, however, that the evaluations given by Seth are not limited to, nor are focused on, the philosophical aspects of consciousness. For instance, Seth allocates considerable space to the neurological and physiological aspects of consciousness, which falls outside our domain of interest and focus.

In evaluating computational models of consciousness from a viewpoint of philosophical scrutiny, the study has reached three major achievements. 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 composite and eclectic 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 the composite and eclectic 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 composite 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, it is often the case that such a unifying framework acts as a facilitator or mental aid in mediating the different views of researchers that make up the constituents of the interdisciplinary study. Take, for instance, the case of phenomenal consciousness, which is considered by some philosophers to be unexplainable or 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 philosophy in revising their comprehension. 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 bedded 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 some certain practical reasons. Firstly, reaching the source codes of some models, like IDA that is developed for the U.S. 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 enough 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, along with the time restrictions we faced, 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.

This study is organized in four chapters. After the current chapter of introduction, we discuss the philosophical aspects of consciousness and develop our own list of features. Chapter 2 starts with a summary of relevant literature of philosophy concerning consciousness. It also reviews three similar lists of features of consciousness reported in the literature. A critique of these lists precedes our proposed list. This list is the first significant outcome and the contribution of the study. Subsequently, in Chapter 3, we start by revisiting the concept and meaning of a model. We also give a brief account of modeling in cognitive science. With all the prerequisites in place we then proceed to evaluate the four self-claimed computer models of consciousness, namely Clarion, IDA, ACT-R and the model based on Cleeremans' review and study (CRS)[2], with respect to our list developed in Chapter 2. These evaluations are conducted in detail, dedicating separate subsections for each of the four computer models studied. These subsections have a similar format: they start with a description of the model and then give an evaluation of the model concerning how well they fulfill the demands of the features in our list. We summarize our findings and state our conclusions in the final chapter, Chapter 4.

2 For brevity throughout this study we shall refer to Cleeremans' review and study by the abbreviation CRS.

CHAPTER 2

A PHILOSOPHICAL OVERVIEW OF THE ELEMENTS OF CONSCIOUSNESS

2.1 Literature Review of Philosophy of Consciousness

By the end of this part, a composite and eclectic list of features of consciousness will be constructed. To this end, we first examine the various lists of features of consciousness taken from several researchers and then construct our own list with an analysis of the items in those lists. However, most of the features in these lists have direct references to the philosophy of consciousness literature without further explanation. So, we found it necessary to give a brief review of the current state of art in philosophy of consciousness. Yet, before presenting this review we think it would be appropriate to introduce some points about the term 'consciousness' as it will be used throughout this study.

There are basically two terms that are needed to be explained in relation to consciousness, namely 'attention' and 'awareness'. What makes attention so related to consciousness is that one can be conscious of only the attended stimuli. In that sense, attention may be taken as “an agency for bringing a stimulus into conscious awareness” (DiGirolamo and Griffin, 2002). Yet, there is evidence that a stimuli that is on the focus of attention may still remain unconsciousness. For example, in the case of blind-sight syndrome, some visual stimuli seem to be attended unconsciously. Blind-sight patients have a blind area in their visual field due to a lesion in their visual cortex. These patients report that they have no visual experience at their blind area. However, experiments reveal that when they

are asked for some information about an object that is at their blind area, their answers are correct by an above-chance percentage.

'Awareness', on the other hand, seems to have a more intimate relation with consciousness. Indeed, the words 'consciousness' and 'awareness' seem to be used as synonyms most of the time in the literature, sometimes accompanied by a note that indicates this equivalence of use without giving further justification. However, there are some works (for example see Chalmers (1996) and Newell (1990)) that postulate a difference between awareness and consciousness. In these works, the motivation behind the distinction is to capture the difference between the phenomenal properties of a conscious experience and information processing properties. Note that, we used these two terms interchangeably throughout this study.[3]

The mind-body problem is an ancient one and 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 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 we could not talk about unconscious mental states. However, as an exception, Leibniz (1997, pp. 53-58) 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. Appercetion can be seen as perceiving that is also accompanied by a knowledge of perceiving. Yet, the state of art remained as equating consciousness with mental states until Freud.

3 Throughout this study, rather than the term 'awareness', we prefer to use the term 'consciousness' as a term to indicate all properties of a conscious experience. The use of the term 'awareness' in this study is due to a faithful adherence to the original use in the literature.

Freud can be considered as the first who conceptualized a framework for unconscious mental states.[4]

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, by the rise of behaviorism, not only consciousness but also any kind of investigations concerning the nature of mental states had been left out of 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 1980's when empirical findings on unconscious processes started to accumulate. This accumulation gave way to treating consciousness as a variable[5] and studying it empirically (McGovern and Baars, 2007).

Correspondingly, there has been also a rise of consciousness studies in the field of philosophy. Besides several philosophical theories that propose some explanation about how mental states become conscious, there is also an extensive literature

4 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 Fruedian framework. The term 'unconscious proper' stands for the mental states or processes that were conscious for sometime but are repressed. The unconscious proper can be made conscious through psychoanalysis. On the other hand, there are preconscious mental states or processes that are only contingently 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)

5 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, contrary to another group who are not conscious of them.

that is devoted to the conceptual clarification of the issue. Blackmore (2004) has the following as the opening sentence of her book: "If you think you have a solution to the problem of consciousness, you haven't understood the problem". However, it is also not clear at all whether there really is "*the* problem of consciousness," since it seems like 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, 1995, 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 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 states “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 the phrase ‘what is it like to be’ such a conscious organism. However, if one tries to understand this aspect, the only thing one may be able to grasp 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 that evidently lacks an explanation of subjectivity.

A similar argument is the “Knowledge Argument” proposed by Jackson. In this case, 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 Mary learns something new: she learns the redness of red or *what it is like* to see red.

Levine (2007) accepts that Marry 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 physical theories that are proposed in the scope of science and subjective experience in a conscious experience.

2.2 Lists of Features of Consciousness

Everything starts with what to call conscious and why to do so. Without clarification of this issue, it seems like the theories we propose are bound to have inconsistencies. Each of us knows something about consciousness. We have some intuitions about it. In any other domain one can keep theorizing without taking intuition into consideration. However, in the case of consciousness, it seems like we all want a theory that gives us an explanation about our intuitions. This is not to say that all our intuitions will turn out to be scientific truths: they may be illusions. However, any theory should at least give us an understanding of the mechanisms underlying these illusions.

With the above considerations, it is 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.

In the recent literature, philosophy has identified a set of features and aspects of consciousness. Some of such lists are already used in the evaluation of various theories of consciousness. However, it should be noted that in most of these evaluations, the features are taken in a somewhat simplified way.[6] Besides these evaluations, there is also a classical paper by Bechtel that should be mentioned here.[7] In his 1995 paper titled "Consciousness: Perspectives from Symbolic and Connectionist AI", Bechtel focuses on the three features of consciousness and investigates their implementation possibilities by two traditions of computational modeling. The three features are "intentionality", "first-person privileged access", and "qualitative character". Intentionality is the aboutness of the conscious mental representations, i.e. what they intend to represent. Bechtel proposes that intentionality can be implemented via the relations between different representations. Privileged access indicates one's having direct access to the contents of one's own conscious mental states. For privileged access, Bechtel claims that this may be indeed an illusion and that all we do is to learn how to express our mental states. Lastly, Bechtel (1995, p.1085) states that "the least tractable feature of consciousness from a computational perspective would seem to be the qualitative character that seems to attach to conscious mental states."

In this section, we first want to explain these features as they originated in the literature. In our review, we will be faithful to the developers of the lists. That is, some of the concepts that are used in the lists may have various different implications in the philosophy of consciousness literature. Nonetheless, we will present the items exactly as they are used by the authors. As a result, this section

6 For example, Sun and Franklin (2007) claim that different representations of knowledge in a computer model may suffice as an explanation for the phenomenal consciousness. However, when one examines the concept and implications of the phenomenal consciousness, one can see that there is more to phenomenal consciousness than representation.

7 Bechtel's paper is a relatively classical one. We find it noteworthy since it is one of the first works that investigates the possibility of implementation of consciousness, with a focus on philosophical features of consciousness.

will be a detailed review of the lists of the features of consciousness. We present our own composite and eclectic list in the next section.

2.2.1 Block

One of the earlier analyses of the features and aspects of consciousness is due to Ned Block (2007b). He differentiates four kinds of consciousness: *phenomenal consciousness*, *access consciousness*, *monitoring consciousness*, and *self-consciousness*. Block regards his analysis to partly aim 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. It can be the case 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 art is considered, the classification, at least at a conceptual level, seems 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.

The first concept of consciousness that Block distinguishes is *phenomenal consciousness*. He starts analyzing this concept by synonyms, i.e. he gives the terms that have similar meaning with 'phenomenal consciousness' rather than giving an exhaustive explanation of that term. According to him one such synonym is the term 'experience'. He states that properties of phenomenal consciousness are the properties that the subject of the experience experiences. That is, they are the properties of experience as they appear to the subject of the experience. What the subject experiences is at the heart of phenomenal consciousness. As Block put it, "what makes a state phenomenally conscious is that there is something 'it is like' to be in that state" (Block, 2007a, p. 275). Block explains phenomenal consciousness mostly with direct reference to the philosophy of consciousness literature and to some hypothetical examples.

Examples are of the kind where one can see the difference between phenomenal and access consciousness.

Access consciousness, on the other hand, comprises the functional properties of a conscious experience. For an experience to be called access conscious, the content of the experience must be available to the direct "rational" use of the subject. The emphasis on the direct rational usability is in order to exclude some cases like blind-sight patients as having access consciousness.

Block does not consider blind-sight patients as access conscious since they cannot report anything about the objects in their blind area unless they are asked. In that sense, although it can be said that these patients have partial access to the information about the objects in their blind area, they cannot do this in a direct rational manner. Nonetheless, Block gives a hypothetical example of the patients that he calls "super blind-sight". In this case, patients somehow learn getting the information without being asked. So, they can report or use the information by themselves. Block proposes that although they do not have phenomenal consciousness, super blind-sight patients have access consciousness, since they can use information in a direct rational way.

In his paper titled "Concepts of Consciousness" Block (2007a) states access consciousness as being the information-processing correlate of phenomenal consciousness. However, in his work (2007b) "On a Confusion about a Function of Consciousness" he also acknowledges that these two can have separate neural correlates.

The other two concepts of consciousness that Block identifies are *monitoring consciousness* and *self-consciousness*. Monitoring consciousness can be seen as a higher-order consciousness. It can be an internal scanning-like mechanism, or a second order mental state that accompanies first order conscious mental states. Block does not propose any particular mechanism or property for this kind. Rather, he points to the inconsistency due to not distinguishing it from other

kinds. An example that he quotes is from Rey (1983, in Block 2007b). Rey states that internal scanning is crucial to consciousness, yet ordinary computer also has a function of scanning. Accordingly, Rey concludes that our understanding of consciousness is incoherent since nobody is eager to call ordinary computers conscious. Now, according to Block, the seeming incoherency is due to not distinguishing the other kinds of consciousness from monitoring consciousness, and by taking monitoring consciousness as if it is all there is to consciousness.

Self-consciousness for Block means to be able to conceptualize a self. He states that all phenomenal consciousness states have a perspective of the subject. Yet, self-consciousness is not identical with this perspective. In order for a being to be self-conscious, having a perspective is not enough. A conceptualization of, and so the knowledge of one's self, is needed for self-consciousness.

2.2.2 Van Gulick

Robert Van Gulick (2005) 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 to talk 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. For example, as indicated above, blind-sight patients have mental states that represent at least some information of the object in their blind spot. However, these mental states are different from the ones that a healthy subject would have. So, an explanation of this difference is needed. The mentioned mental states of blind-sight patients are unconscious mental states. They are potentially conscious in the sense that they would be conscious mental states if the patients did not have the

blind-sight syndrome. Yet, there may also be nonconscious states, in the sense that there is not any possibility of them to be conscious.

Another distinction that we want to understand about consciousness is the difference between conscious and unconscious creatures. *Creature consciousness* may be seen as a property that is ascribed to some creatures.[8] It is the creature consciousness that one talks about when stating that an amoeba may not be conscious at all. It seems plausible to say that having a conscious mental state at some particular time is a necessary prerequisite condition to calling that being creature conscious. However, Van Gulick rejects this idea. He states that, for example, if we accept a particular theory of consciousness, it would turn out to be unfair to expect a fish to have conscious mental states in the sense that the theory proposes. Nonetheless, according to Van Gulick, we cannot call it unconscious. He says that this may be setting the standard somewhat too high.

Also, an important feature of consciousness, according to Van Gulick, is our being able to use its content directly. That is, we have knowledge about the content of a conscious mental sate. As Van Gulick (2005, p.65) puts it, conscious mental states "have meaning or content *for* the person or creature whose states they are."

The last three features – namely qualia, phenomenal structure and subjectivity – of the list have a common point. Van Gulick states that these three features are usually referred to as the phenomenal aspect 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.

The concept of qualia indicates the subjective qualitative properties of an experience. They are also sometimes called the “raw feels” of an experience.[9]

8 We can distinguish the two senses of consciousness as *creature consciousness* and *state consciousness*. Creature consciousness denotes the property of a creature that is capable of being conscious as explained above. On the other hand, state consciousness is a property of mental states, in the sense that a creature's mental states can be conscious or unconscious.

9 Consider, for example, a person who tastes mastika liquor (μαστίχα) for the first time. Even if the person knows the taste of mastika gum and of alcohol, there is something new about the experience of tasting the liquor for the first time. That which is not available to the experiencer before drinking the liquor and that becomes available after drinking the liquor is the quale

The exact nature, or even the presence, of such qualities is problematical. However, all conscious experiences, and only the conscious experiences, seem to be accompanied by a raw feel. So, we must either understand what qualia are, or why we seem to have such raw feels.

Moreover, our conscious experiences have a unity. That is, for example, despite the various modalities of sensations that exist in a conscious experience, we always perceive this diversity as a unified whole. Van Gulick states that this unity is due to a conceptual structure of our conscious experiences. This structure, he calls *phenomenal structure of experience*.

The last feature, which Van Gulick proposes, is the subjectivity of conscious experiences. He states that some aspects of conscious experiences can be known only by the subject of the experience, or by the ones who have similar experiences. This knowledge is a first-person point of view knowledge and cannot be attained by third-person explanations. Van Gulick gives the examples of knowledge of the bat in Nagel (2007) and knowledge of Mary in Jackson (2007).

2.2.3 Lycan

William Lycan (1999) points to the confusion on the consciousness literature. 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 states is the difference between conscious and unconscious states. Accordingly, one thing that we must discover is, how and based on what some mental states are conscious and others are not.

---

associated with the taste of mastika liquor. Also, when Marry sees red for the first time, as explained in the previous section, she becomes acquainted with the quale of red.

Secondly, one seems to have a knowledge of the content of one's conscious experiences via introspection. Yet, this introspective knowledge is not readily available to any other person. This process of introspection also needs an explanation.

Another problem is the concept of qualia. The concept of qualia, in the sense that it is explained above, takes its place among the problems Lycan cites as a problem that must be investigated separately.

Conscious experiences have a 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 materials, our experience of these properties is smooth and continuous. This homogeneity must be also explained.

Yet another feature is the *intrinsic perspectivalness* of conscious experiences, i.e. 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.

The other three problems are the *funny facts*, *the ineffability of "what it's like"*, and *the explanatory gap*, as Lycan calls them. Lycan 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. Lastly, Lycan points to the "explanatory gap" problem that Levine (1997) raises. Lycan states that the connection between the physical, i.e. neurological, facts and the subjective properties of a conscious experience must be explained.

2.3 A Composite and Eclectic List of the Features of Consciousness

Now we have all the sufficient background information upon which to construct our composite and eclectic list of features of consciousness. This list is derived on

the basis of the three authors' lists that were examined in the previous chapter. Most of the features of consciousness that are identified by the above lists are similar. So, it suffices to group the similar features that appear in different lists as a single element in our list. Secondly, we omit some of the features given by some authors. We will first present our list. The explanation as to why such elements were omitted in our list is defended towards the end of this section.

The order of the items in the list does not mean to indicate any hierarchy. However, as it is explained in the following paragraph, the first item, namely difference between conscious and unconscious mental states, indeed has a special place with respect to any theory of consciousness. Also, the last three items, i.e. qualia, phenomenal structure, and subjectivity, share a relevance with phenomenal consciousness. As such, they appear as a contiguous block of items. Apart from these, there is no further meaning to the order of the items. It is especially important not to attribute any hierarchy of item importance to the order, since one of the basic claims of this study is that any theory of consciousness should be able to explain at least all of the items in the list.

The first element in our list concerns the difference between conscious and unconscious mental states. That is, a theory of consciousness is necessary to explain how some mental states become conscious and others do not. This item is different from all others in the sense that it is the most obvious aspect to be explained. In fact, this first element would be a good starting point in the establishment of any comprehensive theory that attempts to explain consciousness. Some features of consciousness may be implicitly omitted from a theory of consciousness. But no theory omitting this first element can be considered as a viable theory of consciousness. This item appears in the two lists that we considered above, namely in the list by Van Gulick and by Lycan. It is not stated in the list of Block. However, it should be noted that Block is, in a sense, 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 similar to what Block called *access consciousness*. That is, the content of a conscious state must be available to the 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 to the concept of availability. However, in both of these notions the emphasis is on the knowledge of the content of the conscious state. This knowledge may be interpreted to correspond to the *monitoring consciousness* in Block's distinction. Accordingly, this difference leads us to the next item on our list.

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 consciousness in the way that Block defines *monitoring consciousness*. Similarly, *introspection* discussed by Lycan, may be seen as such a higher order process. Introspection yields explicit knowledge of the content of the conscious mental states.

We include our list the concept of qualia as a feature of consciousness. It is clear that we seem to have some subjective raw feeling 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 feelings 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. That is, our conscious experiences seem to have a unity. 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 it is stated both in the list of Lycan and of Van Gulick (as subjectivity and the intrinsic perspectivalness, respectively), our conscious experiences have an essential point of view of the subject.

It is worth to note here that we do not propose any dichotomy regarding the features of consciousness. Rather what is proposed is that the conscious and unconscious mental states have differences regarding the features in our list, and there is no obstacle for these differences to accept a grading. If one thinks that consciousness has degrees or levels, (as most of us would), we can easily account for that by referring to the fact that some of the items in our list, such as difference and availability, also admit of degrees or levels.

The six features of consciousness that we identified with inspiration from the philosophy of consciousness literature to date constitute our compiled list. Below is a summary that charts the features of consciousness in our list.

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 content of conscious mental states, monitoring consciousness
4	Qualia	Raw feels or sense data that make up the "qualitative features" of an experience
5	Phenomenal Structure / Unity	Different modalities of perception 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.

As it is seen, we omit some items from the lists given by the three prominent authors in the current literature. So, we will now give our reasons for those omissions. But before that, strictly speaking we do not omit phenomenal consciousness on Block's list. Rather, we follow Van Gulick and divide 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 Block identifies it.

The first item that does not appear in our list is *self-consciousness*. Subjectivity has surely something to do with the self. However, self-consciousness, as it is defined by Block, is different from subjectivity. As it appears in Block's list, self-consciousness requires to have a concept of self and to be able to use this concept. We think that the concept of self is a different issue from 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 thus insufficient to fully present self-consciousness without an accompanying theory of the concept of self.

Secondly, we also omit *creature consciousness* in the list of Van Gulick. As it is mentioned earlier, Van Gulick states that it may be unrealistic to expect some animals to have a specific kind of mental states, but that still we may want to call this animal conscious. That is, proposing that a creature is conscious only if it is able to have some conscious mental states, may be to set the standard too high. Accordingly, he opposes the idea to explain creature consciousness in terms of state consciousness and takes creature consciousness as a separate item in his list. However, if a theory of consciousness is ever developed, then we must bite the proverbial bullet and accept the standards of the theory. So, we think that creature consciousness is explicable in terms of a creature's having state consciousness.

Further, *funny facts*, *the ineffability of "what it's like"*, and *the explanatory gap* in the list of Lycan 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. On the other hand, these arguments are based on the concept of qualia and the subjectivity of conscious experience. So, it is plausible to think that if our understanding of the concept of qualia and subjectivity improves, we will also be able to explain *funny facts*, t*he 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 from the concept of qualia and from the notion of subjectivity.

CHAPTER 3

COMPUTER MODELS OF CONSCIOUSNESS

3.1 What is a Model?

Prior to our study of computer models of consciousness, it will be useful to first discuss what is actually meant by a model and by the effort of modeling. We 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. In these cases, both 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 such different usages of the term 'model'. However we must squarely address the different interpretations of the terms 'model' and 'modeling' from a philosophical viewpoint. The most important point in the scope of this study is to understand of the relationships between a model and a theory.

There are many types of models that can be considered as being scientific. Mathematical models or formulations, say a nonlinear optimization model encountered in microeconomics, or 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 features of the model and the aspects of the phenomena being modeled. In fact, in physical iconic models the model is often a scaled down version of the actual thing. 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 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 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 is established as indisputable explanations. Accordingly, many computer models of consciousness are used as investigative tools to assist in the acquisition of the insights that may potentially lead to theories of consciousness as opposed to mechanically implement well established and generally accepted theories of consciousness. In fact, at this time there are no well established and widely 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 and French, 2000). The Received View considers the model to be often simplified illustration of the theory. This simplification may be valuable in explaining the theory or illustrate its various aspects. A model of a water molecule may be constructed by two small and one large pingpong balls connected by wire. The Received View proposes that all scientific models are illustrations and simplifications as in this example. This view implies that a theory for the phenomena exits, and the model simply illustrates this theory. The theory-model relationship is one-way. The model does not contribute to the theory. It simply illustrates it. In this sense, the model explains the theory using different elements. It is akin to translating a given concept from one language to another.

The Semantic view corresponds to a slightly different view of what a theory is. It has been suggested that a theory provides the structure by which our knowledge

of a given phenomenon may be organized. With this general view, a 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 that describes that theory. Newton's Theory of Gravity may be summarized, for instance, by his well-known formula. Equally so, the theory may 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 has also been interpreted as the theory being a set of all possible models that could be realized from the prescribed 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 nonetheless attempts to construct computational models to gain insights into the behavior of the underlying complex and often dynamical systems. For example, as one may attempt to construct a computer model of animal migration patterns, one would need to add descriptive relationships among the system constituents and assign values to the pertinent parameters. 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 in parallel formulating a tentative theory. 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.

With the brief classification given above, we see that computational models of consciousness would best be considered as being examples of the Autonomous View. As it will be discussed in the following sections, these computational models proposed and developed structures and parameters as the needs so arise to achieve an operational and executable model. In this sense the computational

models could be viewed as proposing structure or theories to explain the various aspects of consciousness.

1960's 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 and Fox, 2002). The first school, mostly called as 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. Besides the pure symbolic approach and connectionist approach there are also hybrid models that attempt to combine the better aspects of the two.

Historically, both these schools were established around the same time. In the early years the neural network approach faced drawbacks because it had been claimed that no linearly weighted neural network could implement Exclusive OR (XOR) function.[10] This obstacle was finally overcome by the introduction of the so-called hidden layers sandwich between the input and output nodes. A second

10 Exclusive OR (XOR) is a logical operation that results true if exactly one of its conditions is true. See, also, Bechtel and Abrahamsen (1991) for a detailed review of XOR problem in connectionist networks.

wind of activity in the neural network approach took place starting in the 1980s. 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. In this sense the neural network is a more holistic instrument while the symbolic rule-based approach tends to be closer to the more customary formal systems.

It was mentioned that the desire to build thinking machines gave impetus to the computer models of cognition. There are differences between the purpose and motivation behind 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 CMattie and IDA 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, and similarly, versions of the scientific models could form the basis of future practical products.

In fact there are many practical software products that take on the responsibilities of humans. For instance, there are many expert systems based products that are very successful in automotive engine diagnostics.[11] However, these products do not claim to be inspired by, or to implement, aspects of human cognition. So they are not as appropriate to be subjects of examination in cognitive science. Models

11 Engine management computers routinely collect and store information on engine performance. Such information is downloaded during routine maintenance to a diagnostics computer that runs an expert system. This enables the identification of any problem with greatly improved efficiency in automotive repair.

that are specificity developed to enhance our understanding of cognition rather than to develop practical products are hence more prevalent in the cognitive science literature.

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 proposed 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 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, would have significant implication as the structure of cognition. The work of Maia and Cleeremans (2005) which combines three task-specific models towards a more general framework for consciousness is a good example of this possibility.

3.2 Computer Models of Consciousness

This section reviews some self-claimed computer models of consciousness that have appeared in the literature. It should be mentioned that there are a relatively small number of such models. Below we study Clarion, IDA, ACT-R and the model proposed in the Cleeremans' review and study.

3.2.1 Clarion

Clarion (Connectionist Learning with Adaptive Rule Induction ON-Line) is a cognitive architecture developed by a team led by Ron Sun (Sun, 2003; Sun, 2006). Clarion is a rather hybrid architecture that involves both connectionist and rule based elements. It can be used to model various cognitive tasks in many domains. Clarion recognizes implicit and explicit knowledge. It uses a back

propagation network to represent subsymbolic implicit knowledge. Semantic rules are used 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.

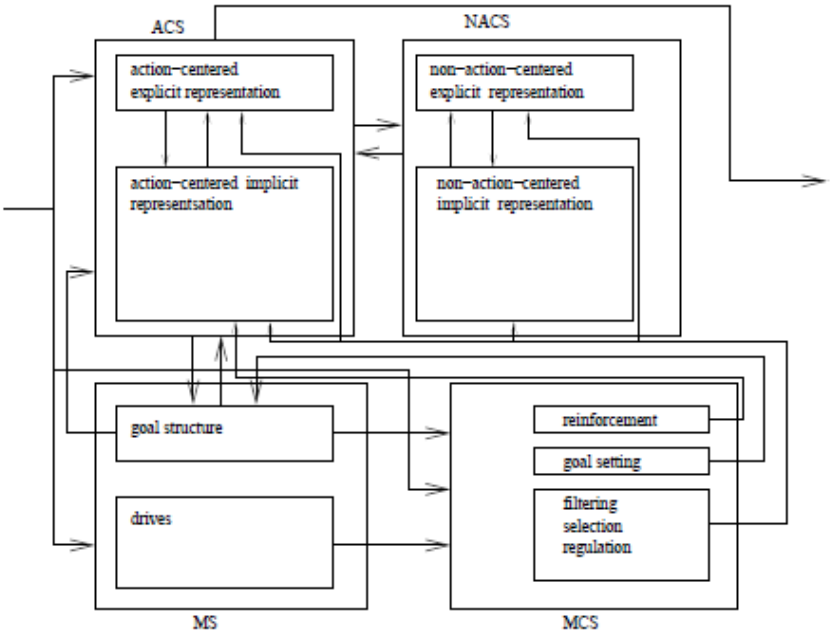

Figure 1. The Clarion Architecture (adapted from Sun (2006))

Explicit knowledge works with rules and chunks. Chunks are collections of dimension/value pairs. Here the word 'dimension' refers to properties. For example, the chunk car-a may consist of the dimension/value pairs, (type, sedan), (color, red), (number_of_doors, 4). An important concept in Clarion is the so called "state of the world". It refers to the environment in which Clarion is to operate. The state of the world consists of many dimension/value pairs. In addition, Clarion implements a working memory[12] as well as goals which also

12 In psychology, "the concept of working memory proposes that a dedicated system maintains and stores information in the short term, and this system underlines human thought processes" (Baddeley, 2003, p. 829). The term 'working memory' was used in the work of Baddeley and Hitch dated 1974 (as cited in Baddeley 2003). The core idea of working memory has been

contain 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 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. That is, both the top level and bottom level use the same dimension/value pairs, although the top level combines dimension/value pairs into chunks. Clarion implements different types of learning for the top and bottom levels. The bottom level, back propagation networks use Q-learning. Q-learning basically considers many alternative actions and rewards the action that is the most helpful towards the set goal. There are many alternate learning mechanisms for 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.

As mentioned, some dimension/value pairs are kept in what is called working memory. Working memory is akin to human short-term memory. Dimension/value pairs are placed in working memory as a result of the outputs from the rules or backpropagation networks. Items placed in working memory degrade and decay over time. This allows the content of the working memory to be changed over time. Note that items in the working memory are available to the rules at the top level and it is also available to the bottom level as input. As a separate component Clarion implements a goal structure and maintains a goal stack to handle the hierarchical nature of goals and sub-goals that must adhere to given precedence relations.

---

preserved from that time, although there have been some modifications to the original model. For a review of the studies on working memory see Baddeley 2003.

We now review Clarion with respect to our composite list. The first item on our list was the difference between conscious and unconscious processes. Clarion has a two-level architecture. The lower level neural network corresponds to 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, availability, does not hold in Clarion. Although information can be shared among the subsystems, no attempt is made to elevate the top-level conscious processes more so than any other information. Lower-level information is shared as much as upper-level information. In other words, conscious thoughts are shared in the same manner as unconscious thoughts. We must thus conclude that Clarion does not fulfill the second item on our list.

The third item on our list, direct and explicit knowledge is not addressed in Clarion. There is a general view that models that employ symbolic representations and rules at the upper-level are more suitable for implementing direct and explicit knowledge. However, just representing such knowledge is not sufficient to fulfill the third item on our list. We would expect the model to not only represent such information but also provide specific mechanisms to actually be able to use and act on such information. Clarion lacks the latter.

The last three items on our list are qualia, structure of experience, and subjectivity. All three items are related to phenomenal consciousness. Our view, as explained in the previous chapters, is that phenomenal consciousness is interpreted in different ways in the literature. This is why we explicitly list qualia, structure of experience, and subjectivity as separate items, although other researchers have used the term 'phenomenal consciousness' to describe 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 structure of experience or subjectivity. What Clarion claims to address may be considered to fall under the item qualia. Although the separate representation of upper and lower level processes seems to be a plausible structure to explain qualia, simply implementing such a division is not, in our view, sufficient to model qualia.

3.2.2 IDA

IDA (Intelligent Distribution Agent) (Franklin, 2000; Franklin, 2003) is somewhat different from the other models in that it has been given a specific practical task to perform. Namely, IDA is used by the U.S. Navy to assign specialists to new posts. This task requires several things. IDA must communicate with the sailors through e-mail. It uses natural language in doing so. IDA must access the databases to see what is needed and what is available. Moreover, IDA must 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 in quotations, following the developers, since there is no claim to model all aspects of consciousness. For instance, it is explicitly stated that IDA has no claim in modeling phenomenal 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 the environment, and acts on it, over time, in pursuit of its own agenda" (Franklin and Graesser, 1997). 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 was put forth by Baars (1988). In that work, it is proposed 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. IDA combines components responsible for so-called high-level abstractions such as behavior and emotions 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. These codelets constitute the multiple agents in the multi-agent view of IDA. Perception, for instance is accomplished by the many codelets acting on the inputs. Each codelet tries to identify some property of the input. 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. In content-addressable memory, the address of the data need not be known. The address is retrieved, at least in principle, from the data. For instance, when given a last name, a human may easily recall the first name, and the profession of the person. No address information is needed, as partial data is sufficient to retrieve the entire contents of memory. In IDA, 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."

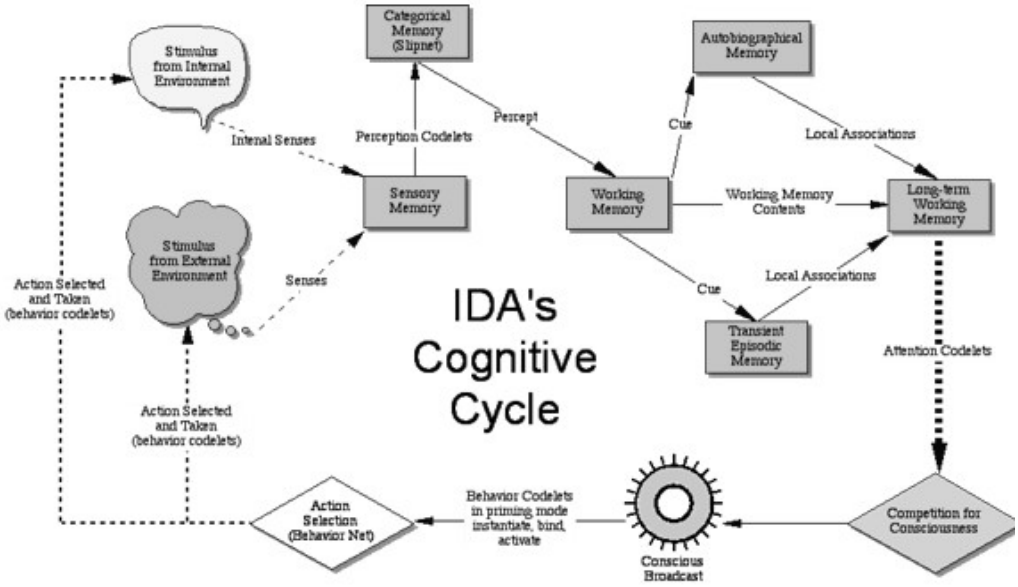

Figure2. IDA's Cognitive Cycle (adapted from Franklin (2009))

As described, consciousness is modeled by the multitude of codelets, some of whom 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. Typically, the coalition with the highest strength is chosen by the spotlight controller. 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.

All this takes place in cyclical manner. The inputs are scanned by independent parallel codelets. Attention codelets operate on the inputs and coalitions are formed to compete to enter the global workspace. Those coalitions chosen by the spotlight controller enter the global workspace and are announced by the broadcast mechanism. The action is generated by the behavior codelets that implement the production rules. The IDA input-processing-output cycle thus parallels the psychological stimulus-cognition-response cycle.

We now review IDA with respect to our composite list. The first item on our list was the difference between conscious and unconscious processes. IDA implements a global workspace. Information that is elevated to the global workspace is considered to represent the content of conscious mental states. In this respect, IDA fulfills the first item on our list.

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

The third item on our list, direct and explicit knowledge is not addressed in IDA. There is no mechanism to maintain a representation of IDA's internal states. This, given the task-oriented nature of IDA, is understandably not taken to be a priority. Nor is there any claim that IDA attempts to do so.

The last three items on our list are qualia, structure of experience, and subjectivity. Again, being task-oriented, IDA 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 is sometimes confused, we find little credence or value in this claim. In all fairness, however, IDA may have the beginnings of some aspects of what we called structure of experience. Specifically, we mentioned the unity of conscious experience. The coalition manager and the

related structures of IDA 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 handle 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).

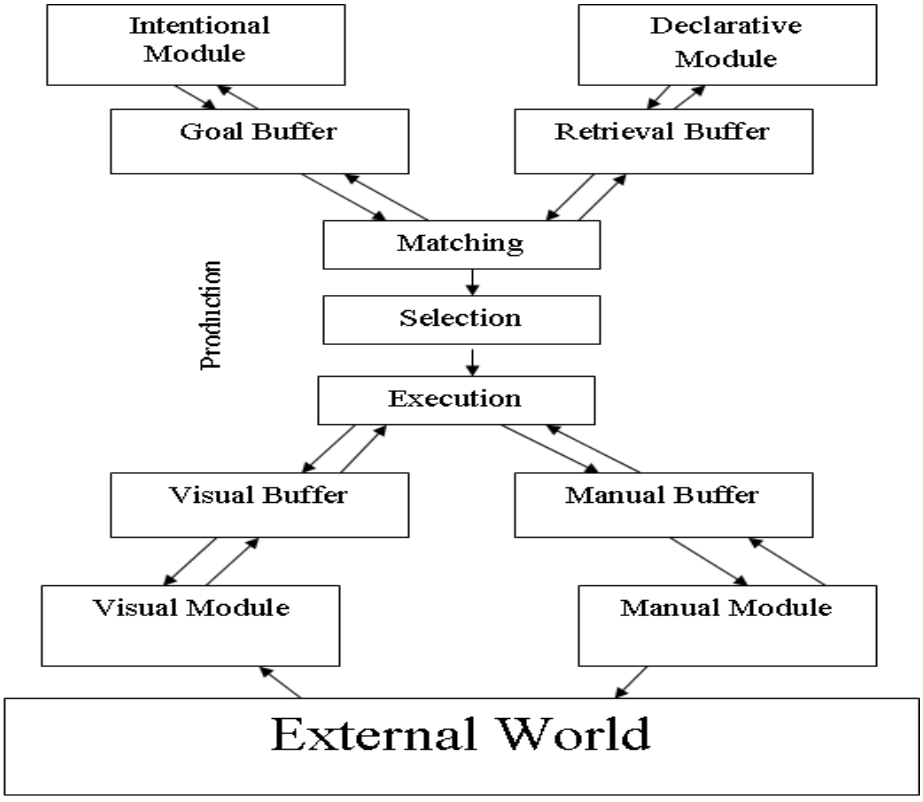

Figure 3. ACT-R Architecture (adapted from Anderson et al. (2003))

ACT-R attempts to integrate several modules with a core production system. The exact number of modules is not important. A set of modules have been implemented and made available. Users may insert their own production rules and experiment with the ACT-R architecture. As an example, consider the implemented vision module, motor module, intention module, and declarative memory module. The vision module has the task of identifying objects in the field of vision. Critical to the concept of a module, each module contains a buffer. Just as a person may see many objects in her field of vision but focus her attention on a specific object, the buffers selectively hold the information most relevant to the task at hand. In fact, buffers play a central role in ACT-R. ACT-R proposes that these buffers interact, with the involvement of the production rules, to determine cognition. Specifically, the production rules scan the buffers. 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 various modules in ACT-R, in general, perform their tasks in parallel, independently of each other. 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.

As mentioned, ACT-R does not make an explicit attempt to model consciousness *per se*. However, that the resultant model displays elements of consciousness. To

start with, not all of the information in the modules are placed in the buffers. The distinction between the information in the modules and in the buffers fulfill 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, direct and explicit 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, structure of experience, and subjectivity all relate to phenomenal consciousness. Again, ACT-R does not address these elements at all.

3.2.4. Cleeremans' Review and Study

The last computational model is based on Cleeremans' review and study, CRS (2005; Maia and Cleeremans, 2005). CRS is different from the previous three. Although the ideas have been proposed, a computer model that exactly implements those ideas has not yet been developed. Nonetheless, we regard the CRS to be in the same category as the models discussed above, since it can, in principle, be implemented. Cleeremans develops his ideas based on observations from three separate computer models. In addition he shows how the final ingredient of his model, namely, the issue of meta-representation, can be implemented.

The first of the models, that is used by Cleeremans is by Dehaene et al. (Dehaene and Changeux, 2005). It models the phenomenon known as attentional blink. Attentional blink refers to the case where two inputs are issued at rapid succession. While the subject focuses his attention on the first task, he misses the second input. In essence, his attention blinks during the second input and causes him to miss the second input entirely. Attentional blink is modeled using neural

networks. In fact 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. Whichever network receives the input first starts its computation ahead of the other and becomes the first to activate 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 other 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.

The second model by O'Reilly, Braver, and Cohen (1997) implements the so-called AX-CPT (AX-Continuous Performance Test) task by a neural network. Here a sequence of characters are presented to the subject. If the character X immediately follows character A, the subject is to press the left button. In all other cases the subject is to press the right button. The model contains several layers. The stimulus contains a dedicated character for each input. As the characters are received sequentially the corresponding input nodes of the stimulus are activated. The response contains two output nodes, one for the left button and one for the right. The input nodes of the stimulus are linked to the output node of the response to an intermediate layer called PMC (posterior perceptual motor cortex.) In addition, two intermediate nodes constitute the higher-level PFC (prefrontal cortex.) The two nodes of the PFC labeled 'left to X' and 'right to X' are activated based on whether the character A is received before the character X or not. In essence the node 'left to X' of the PFC is activated when the first character A is received. This arms the system. If a character X is received immediately after the system is armed, then the subject is to press the left button. If a character other than X is received, the system is disarmed. With this scheme, only when the character X immediately follows a character A, will the system be armed and the left button pressed. The AX-CPT model uses the PFC to influence the processing that takes place within the PMC and allows the neural network to operate as the human subjects do. That is, there is competition between the actions of pressing the left button and pressing the right button, and this competition is

not only a result of the input stimulus but also controlled by the state of the PFC (whether the PFC is armed or not). Here the PFC corresponds to the working memory and stores the information about the last received character. As Cleeremans and Maia (2005) points out "the PFC layer is capable of biasing competition 'downstream' (via its projections to PMC) and maintaining activation during a delay".

The last neural network model is developed by Cleeremans, Timmermans and Pasquali (2007). It illustrates the the phenomenon of meta-representation. The immediate task is to recognize a single digit given in a five-by-four dot matrix representation. The dot matrix contains a total of twenty input nodes, different combinations of which are activated depending on which number is to be shown. The output consists of ten nodes, representing the digits from zero to nine, only one of which is activated depending on the input. The input and output layers are linked through a hidden layer containing either five or ten nodes. Such neural networks can be trained to successfully yield the correct number based on the input pattern. Cleeremans adds to this structure another sub-network whose outputs contain representations of both input and output nodes of the original network. That is, the subsystem contains thirty output nodes twenty of which are expected to follow the dot matrix input and ten of which the state of the output. In essence the outputs of the subsystem represent the state of the original system. The subsystem receives its input from the hidden layer of the original network through an additional hidden layer of its own. When the entire system operates as intended the network identifies the dot matrix pattern and generates the correct output, while the subsystem reproduces the state of the system from information through the hidden layer. The output of the subsystem corresponds to the meta-representation of the system. This phenomenon is significant since it is analogous to being aware of one's own thought.

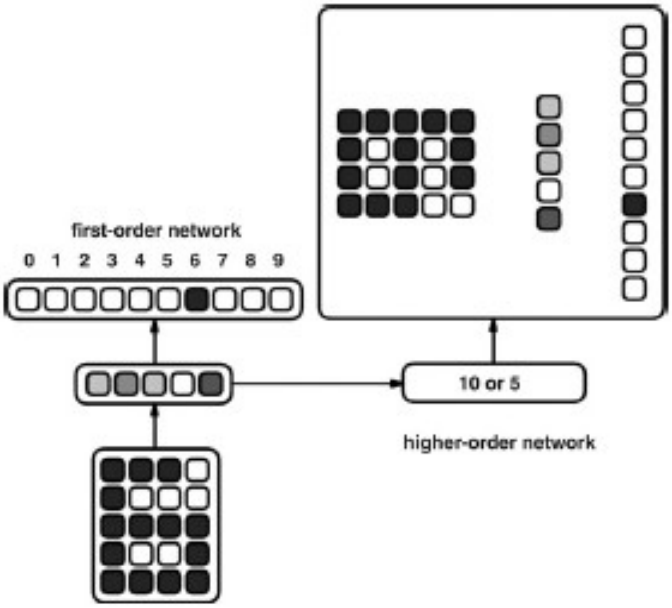

Figure 4. The Meta-representation Network (adapted from Cleeremans et al. (2007))

Based on these four studies, Cleeremans put forward a computer model that may in principle be implemented. He hypothesizes that the following five mechanisms are associated with consciousness, and thus should be included in the model. These mechanisms are, active representation (neuronal firing), global computation biased by top-down modulation (as in the AX-CPT model), global constraint satisfaction (related to the brain's knowledge of the current situation), re-entrant processing (to allow global interpretations in higher level areas to influence the processing in lower level areas), and finally meta-representation (the knowledge of one's thoughts).

The distinction between conscious and unconscious representations is considered by Cleeremans to be dependent on whether the information possesses the qualities of strength, distinctiveness, and stability. The strength corresponds to the level of activation. The net with the highest activation level is the strongest. The net that is selected to be reinforced by the top level network receives its distinctiveness. For instance, in the AX-CPT experiment, once the system is armed by the detection of an A, the pattern that represents X becomes distinctive. Stability refers to the dwell of the state for some time. This is achieved by re-entrant

signals and networks that keep the activation of some states on hand. This dwelling allows time delays between the A and the X, as well as provides the necessary mechanisms in modeling attentional blink. In the latter experiment, the first stimulus is retained by re-entrant signals, thereby inhibiting the process of the second stimulus.

Availability, the second item on our list, is not explicitly modeled. However, the fact that some information is elevated to the top level, and that this information can influence the activities at the bottom levels is, in principle, akin to the notion of availability. In the AX-CPT model, once the system is armed after the detection of the character A, this information becomes available to the input layer, and influences the subsequent processing of the next received character.

The third item, direct and explicit knowledge of the contents of the conscious states is considered by Cleeremans. In the dot-matrix digit recognition simulation, the system employs a separate network to model its internal representations. This amounts to carrying explicit knowledge about the contents of the systems.

The remaining three items on our list were qualia, structure of experience, and subjectivity. The last two, namely structure of experience and subjectivity, are not addressed by the models. Cleeremans claims to have made inroads into qualia by stating that a quale is an acquired quality through experience. Since neural networks are capable of learning, and formulating the relationships among the objects of representations, qualia will eventually be captured by the model. As will be discussed in the following chapter, although we feel that this argument has merit, as far as our list and our evaluation of computer models of consciousness are concerned, we must consider CRS not to have captured qualia in its entirety.

CHAPTER 4

CONCLUSION

We now summarize the findings of the former chapters. We evaluated four computer models of consciousness according to our list of six features identified by philosophy. The following table summarizes the results. In the table, the features each model successfully implements are denoted by a plus (+). The table entries that correspond to the features not addressed by the respective model are left blank. There are a few table entries, denoted by question marks (?). These correspond to partial models or efforts that may need further amplification.

Table 2 Results of the Study

	Clarion	**IDA**	**ACT-R**	**CRS**
Difference	+	+	+	+
Availability		+	+	
Explicit and Direct Knowledge				+
Qualia	?	?		?
Phenomenal Structure / Unity		?		
Subjectivity / Subjective Point of View				

When Table 2 is viewed columnwise, it reveals information about how successful the corresponding model is in satisfying the features of our list. Here, we see that IDA and CRS fair better. When viewed rowwise, Table 2 shows how well the current computer models address each feature of our list. In this respect, difference and availability seem to be addressed more than the other features. By contrast, features such as phenomenal structure and subjectivity are not fully implemented by these models. Recall that the less represented features are related to phenomenal consciousness.

The presence of many question marks (?) in Table 2 also conveys an important message. Although the models attempt to address these features, there is room for improvement ,in our view, before we can accept that the models completely and successfully address and implement these aspects. Our observation earlier that IDA and CRS implement more features includes our optimistically including the question marks in the final score of the features each implements.

Moreover, the abundance of the question marks and empty cells shows, in our view, opportunities for future work towards closing the gap between computer models and philosophical studies. In fact, which of the four 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 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, at the moment, there are a precious few models reported in the literature. It is plausible to suppose that as the number of computer models increase by one or two orders of magnitude, our insights and understanding will also be enhanced. Some of the computer models contain 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 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 distinction of 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. In 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.[13] 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 "catch all" 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 cooperations should aim to carefully identify the components and elements of phenomenal consciousness so that different disciplines have the same concept of these terms. The identification of the confusion surrounding phenomenal consciousness is considered to be a prominently 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, subjectivity, although a most important feature identified by philosophy, is not addressed by any 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 to be unattainable.

13 In fact, it may be said that it is better to keep the three items together since most of the question marks appear in the corresponding cells. However, we think otherwise. That is, the question marks indeed indicate that we need a more careful analysis of these three items. 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.

As a final note, we want to re-emphasize the point made earlier in the introduction of this study. 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.

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.

Baars, B. (1986). *The Cognitive Revolution in Psychology*. New York: The Guilford Press

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Retrieved from: http://vesicle.nsi.edu/users/baars/BaarsConsciousnessBook1988/

Baddeley, A. D. (2003). Looking Back and Looking Forward. *Nature Reviews Neuroscience*, 4, 829-839.

Bechtel, W. (1995). Consciousness: Perspectives from Symbolic and Connectionist AI. *Neuropsychologia*, 33, 1075-1086.

Bechtel, W. Abrahamsen, A. (1991). *Connectionism and the Mind. An Introduction to Parallel Processing in Networks*. Malden: Blackwell. pp. 1-20.

Blackmore, S. (2004). *Consciousness: An Introduction*. New York: OUP.

Block, N. (2007). *Consciousness, Functionalism and Representation (Collected Papers; v.1)*. Cambridge: MIT Press.

Block, N. (2007a). Concepts of Consciousness. In Block 2007, 275-296.

Block, N. (2007b). On a Confusion about a Function of Consciousness. In Block 2007, 159-213.

Block, N., Flanagan, O. & Güzeldere, G. (1997). *The Nature of Consciousness: Philosophical Debates. Cambridge: MIT Press.*

Chalmers, D. J. (1996) *The Conscious Mind*. New York: OUP.

Clowes, R., and Seth, A.K. (2008). Axioms, Properties, and Criteria: Roles for Synthesis in the Science of Consciousness. *Artificial Intelligence in Medicine,* 44, 94-104.

Cleeremans, A. (2005). "Computational Correlates of Consciousness". In S. Laureys (ed.), *Progress in Brain Research*, 150, 81-98.

Cleeremans, A., Timmermans, B. & Pasquali, A. (2007). Consciousness and Metarepresentations: A Computational Sketch. *Neural Networks*, 20, 1032-1039.

Cooper, R & Fox, J. (2002). Modelling Cognition. In Cooper, R(ed.), *Modeling High-Level Cognitive Processes*. Lawrence-Erlbaum Assoc.

Coward, L. A. and Sun, R. (2004). Some Criteria for an Effective Scientific Theory of Consciousness and Examples of Preliminary Attempts at Such a Theory. *Consciousness and Cognition.,* 13, 268-301.

Da Costa, N. and French, S. (2000). Models, Theories, and Structures: Thirty Years On. *Philosophy of Science,* 67, 116-127.

Dehaene S and Changeux, J-P. (2005). Ongoing Spontaneous Activity Controls Access to Consciousness: A Neuronal Model for Inattentional Blindness. PLoS Biol, 3, 5, 910-927.

Descartes, R. (1973). *The Principles of Philosophy.* Translated by E. Haldane and G. Ross. Cambridge: CUP. (Original Work Published 1644).

DiGirolamo, G. J., and Griffin, H. J. (2002). Consciousness and Attention. In L. Nadel (ed.), *Encyclopedia of Cognitive Science*. New York: Wiley.

Franklin, S. (2009). How the IDA Model Supports the Crick and Koch Framework. Retrieved from: *csrg.cs.memphis.edu/csrg/assets/papers/IDAonFramworikJournal.doc*

Franklin, S. (2000). A "Consciousness" Based Architecture for a Functioning Mind. In A. Sloman (Ed.), Proceedings of the Symposium on Designing a Functioning Mind
Retrieved from: 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, and K. Stenning (Eds.), CogSci2001: Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Mahwah: Lawrence Erlbaum Associates.

Güzeldere, G. (1997). The Many Faces of Consciousness: A Field Guide. In Block et al., 1997, 1-67.

Jackson, F. (1997). What Mary Didn't Know. In Block et al., 1997, 567-571.

Leibniz, G. W. (1997). *New Essays on Human Understanding.* Translated by P.Remnant and J.Bennett. Cambridge: CUP. (Original Work Published 1765).

Levine, J. (1997). On Leaving Out What It's Like. In Block et al., 1997, 543-556.

Lycan, W.C. (1999). Plurality of Consciousness.
Retrieved from: http://www.unc.edu/~ujanel/CogThs.html

Maia, T.V. and Cleeremans, A. (2005). Consciousness: Converging Insights From Connectionist Modeling and Neuroscience. *Trends in Cognitive Sciences,* 9, 397-404.

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: CUP.

Nagel, T. (1997). What It is Like to be a Bat? In Block et al. 1997, 519-528.

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.

O'Reilly, R. C., Braver, T. S. & Cohen, J. D. (1997). A Biologically-Based Computational Model of Working Memory. Retrieved from: http://psych.colorado.edu/~oreilly/papers/OReillyBraverCohen99.pdf

Seth, A. K. (2009). Explanatory Correlates of Consciousness: Theoretical and Computational Challenges. *Cognitive Computation,* 1, 50-63.

Sun, R. (2003). *A Tutorial on CLARION 5.0.*
Retrieved from: 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: CUP.

Sun, R. & Franklin, S. (2007). Computational Models of Consciousness: A Taxonomy and Some Examples. In P. D. Zelazo and M. Moscovitch (eds.), *Cambridge handbook of consciousness.* Cambridge: CUP.

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

Van Gulick, R. (1995). What Would Count As Explaining Consciousness?. In T. Metzinger (ed), *Conscious Experience*. Paderborn: Ferdinand Schöningh.