STRATEGIES FOR MEASURING MACHINE CONSCIOUSNESS

RAÚL ARRABALES, AGAPITO LEDEZMA and ARACELI SANCHIS

Computer Science Department,
Carlos III University of Madrid,
28911 Leganés, Madrid, Spain
rarrabal@inf.uc3m.es

The accurate measurement of the level of consciousness of a creature remains a major scientific challenge, nevertheless a number of new accounts that attempt to address this problem have been proposed recently. In this paper we analyze the principles of these new measures of consciousness along with other classical approaches focusing on their applicability to Machine Consciousness (MC). Furthermore, we propose a set of requirements of what we think a suitable measure for MC should be, discussing the associated theoretical and practical issues. Using the proposed requirements as a framework for the design of an integrative measure of consciousness, we explore the possibility of designing such a measure in the context of current state of the art in consciousness studies.

Keywords: Measures of consciousness; consciousness metrics; hard problem; easy problems.

1. Introduction

Having suitable tools for comparative analysis and evaluation is a key factor for the progress of any new scientific research. Specifically, in young and emerging fields, like the case of MC research, the availability of these sorts of tools is essential. As pointed out by [Seth *et al.*, 2008], the use of such measuring tools could not only help us to assess the progress actually being achieved, but also to indicate what the most promising research lines are. Although a number of authors have proposed different approaches, defining accurate metrics for assessing the level of consciousness of either biological organisms or artificial implementations remains an open problem. One of the main difficulties is related to the characterization of the term consciousness, which can be described from different perspectives. For instance, from the point of view of phenomenology [Carruthers, 2000], consciousness could be measured in terms of the degree of the vividness of conscious experiences. However, from the point of view of access [Baars, 2002], consciousness could be measured in terms of the contents of the mind available for explicit processing. Additionally, the fact that different

theories try to account for consciousness using different (and to a certain extent incompatible) views [Atkinson *et al.*, 2000], leads to the definition of measures that are only valid in the context of the specific theory they support. Although current theories propose totally different explanations for the production of phenomenal consciousness, we think there are many common denominators about other aspects of consciousness across major theories, and this could help us to define interim measures characterized by the current common agreement of what early MC implementations are expected to be. In other words, the main question we analyze in this paper is: can we identify a minimum consensus reached in the scientific study of consciousness that could be pragmatically used to define an (interim) integrative and mostly agreed measure suitable for MC implementations?

It is important to note that an affirmative answer to the former question does not imply in any way that such a measure would be a complete one. Fully satisfactory measures will be possible only when a final "grand unified theory of consciousness" is developed. Nevertheless, we think that adopting this approach can contribute to a reiterative redefinition of better measures that incrementally integrate current scientific insight about consciousness. This exercise will in turn help to evaluate the validity of the hypotheses being considered in the measuring process, thus providing valuable feedback to the quest for a scientific explanation of consciousness. For instance, if according to particular measure there exist cases in which high consciousness scores are assigned to MC implementations that do not practically show the expected conscious behavior, the underlying hypotheses would need to be revised. Essentially, we suggest that, in the specific field of MC, more effort should be put into the development of measures oriented to the "easy problems" of consciousness [Chalmers, 1995], rather than focusing exclusively in the application of different speculative measures related to the "hard problem" [Chalmers, 1995]. We believe that, adopting an evolutionary inspired approach and extrapolating it to the development of artificial conscious machines, the so-called easy problems of consciousness have to be solved in the first place in order to be in a better position to effectively study the hard problem in artificial cognitive systems. Indeed, the current idea of the hard problem could change drastically when all easy problems are solved [Dennett, 1996]. Although phenomenal states can be present in humans in the absence of directly associated behavioral outcome (for instance, during dreaming or daydreaming), the early development of consciousness is rooted in a direct and adaptive interaction of the body with the environment [Humphrey, 1999]. Phenomenal states without associated adaptive behaviors do not make any sense unless the subject is endowed with cognitive capabilities in the first place. Consequently, assuming that analogous developmental principles apply to MC [Arrabales *et al.*, 2009], a strategy directed to the design of machines able to develop phenomenal states without solving first the easy problem does not seem to be a promising approach. Therefore, the effort in the development of measures of consciousness suitable for MC should be initially more focused on the assessment of the cognitive

capabilities associated with consciousness. Whether or not the development of such measures can also contribute to the detection of phenomenal states in machines remain to be seen. The development of MC implementations able to deal with the easy problems will likely shed light on how artificial *qualia* could be produced, and therefore provides new insights about how phenomenal states can exist in machines. This feedback could be used then to define better integrative measures that also consider the generation of phenomenal states in the machine. Adopting this position does not mean that only cognitive measures should be developed now, neglecting phenomenal approaches to consciousness. What we claim is that measures of phenomenal consciousness alone, without considering the cognitive dimension, seem not to be of practical application in the short term. Considering the hypothesis that phenomenal consciousness and access consciousness will develop together in MC implementations, as seemed to happen in biological organisms, a measure integrating both aspects would be much more significant.

In the following we identify the main requirements of a suitable measure for MC discussing the associated problems; then we briefly review the most salient accounts proposed as measures of consciousness, analyzing the issues related with their potential practical use in the field of MC.

2. Designing a Machine Consciousness Measure

Setting aside the discussion about what theories of consciousness are closer to the reality and whether or not they can also be applied to MC, at this point we should identify practical issues that need to be addressed about the theories and the application of associated measures. In this section we aim to characterize the measures of consciousness that could be considered in the domain of MC and we review the practical requirements that a compelling measure for MC should fulfill.

2.1. *Measuring consciousness*

Before analyzing the specific requirements for a MC suitable measure, it is important to consider the main factors involved in the problem of measuring consciousness as typically applied to humans. First of all, a distinction should be made between the concepts of testing for the presence of consciousness versus measuring the level of consciousness. Although considering consciousness as an on/off property can be of practical use in some every day contexts, a rigorous scientific account must be pursued in order to effectively determine a fairly accurate level of consciousness of either biological organisms or artificial systems. This graduation of consciousness could be applied both to creature consciousness and to state consciousness [Manson, 2000]. In other words, the overall level of consciousness of a subject could be assessed in terms of the particular level of consciousness of the mental states he or she possesses. Therefore, a creature not having any conscious mental states at all is considered completely unconscious.

In addition to the level of consciousness of a given mental state, the related explicit content being consciously perceived could also be assessed. Indeed, the explicit content also determines the functionality of consciousness [Seth, 2007]. Along the lines of the argumentation discussed in the introduction section, the cognitive abilities of an agent determine the specific mental content that will be available to conscious states. In terms of the Global Workspace Theory (GWT) the conscious contents would be those gaining access to the working memory, right under the spotlight of attention [Baars, 1988]. The higher the degree of richness and elaboration of these conscious contents is, the higher the potential functionality of the associated conscious experience will be. Multimodality is also a typical feature of conscious experience, i.e., different sensory modalities, like hearing, seeing, and smelling are bound together giving place to an integrated percept. Understanding how different sensory modalities are unified in conscious scenes is known as the binding problem [Revonsuo and Newman, 1999]. The binding capacity of an artificial mind could also be assessed.

Another important aspect to take into account in the definition of a measure of consciousness is the required multidimensionality. Consciousness is a "cluster" or composed property [Sloman, 2002; Block, 1995], and it cannot be measured the same way as simple properties like distance or mass using single well-defined units (e.g., meters or kilograms). A comprehensive measure of consciousness has to take into consideration a set of capabilities and qualities supported by the system and assess how well they are integrated. One example of the multiple facets that can be associated with consciousness is the list of cognitive skills proposed in the scale *ConsScale* [Arrabales *et al.*, in press]. Obviously, a single score could be calculated as an indicator of the level of integration between different capabilities. Nevertheless, this indicator alone would not provide a sufficient characterization of the level of consciousness.

A scientific measure of consciousness has to be, of necessity, a third person approach; however, consciousness is inherently a first person phenomenon. Therefore, approaches exclusively based on behavior assessment can only be considered as an indirect source of evidence of consciousness. In the domain of MC we believe that the first person problem can be circumvented by combining first and third person approaches as suggested by [Dennett, 1991]. In general, the combination of behavioral and non-behavioral (e.g., dynamical complexity [Seth *et al.*, 2008]) measures is required to fully characterize the level of consciousness of a subject.

An additional strategy for the detection of consciousness is the identification of correlates or hallmarks. A number of properties have been appointed as hallmarks of consciousness [Arrabales *et al.*, in press; Edelman *et al.*, 2005; Seth *et al.*, 2005], however they characterize specific levels of consciousness (like accurate verbal report which is characteristic of human-level consciousness) or specific underlying mechanisms for consciousness (like neuroanatomical properties of mammalian nervous systems). If these hallmarks are to be used in an evaluation process they have to be arranged in specific levels [Arrabales *et al.*, in press].

2.2. *Requirements of a suitable measure for machine consciousness*

In the former section we have identified several facts about consciousness that should be taken into account in the design of a comprehensive measure:

- **F1.** Consciousness is a graded property (a continuum rather than a binary property).
- **F2.** A creature is conscious in virtue of its conscious mental states.
- **F3.** Conscious mental content determines the functionality of consciousness.
- **F4.** Conscious content is multimodal, integrated, and differentiated.
- **F5.** Consciousness is a complex multidimensional property.
- **F6.** Scientific study of consciousness calls for the combination of first and third person approaches.
- **F7.** Different hallmarks of consciousness can be associated with different levels of consciousness and different species or machines.

Although this list of facts is not comprehensive, neither free of controversy, we think it reasonably describes the *explananda* of any theory of consciousness as identified by a significant part of the scientific community, e.g., see [Seth *et al.*, 2008; Manson, 2000; Sloman, 2002; Dennett, 1991; Edelman *et al.*, 2005; Dennett, 1997]. Therefore all these aspects of consciousness should be addressed by an integrative measure applicable to MC. Clearly, some of the former claims are still important sources of controversy, and even the completeness of the list is doubtful. However, we believe that, in order to be practical from the engineering perspective, adopting such a somewhat reductionist position in the domain of MC would be helpful, at least until significant results are obtained that force a revision of the active research lines (either modifying existing claims or adding new ones).

In addition to the former considerations, evaluating artificial systems implies further requirements about design, procedures, and applicability:

- **R1.** The measure should be applicable to any MC implementation, independently of the underlying substrate and technology used in the artificial organism.
- **R2.** The measure should be problem domain independent; i.e., applicable to any MC implementation independently of its application domain.
- **R3.** The measure should be computable in a reasonable time using currently available computational power.
- **R4.** The measure should provide qualitative and quantitative characterization of the level of consciousness of the artificial organism (i.e., able to assess graded consciousness).
- **R5.** The measure should provide a multidimensional characterization of the consciousness level of the subject. Given the complex nature of consciousness, a single aggregated score would not be enough to characterize the level of consciousness of a MC implementation (scores exclusively aimed at, for

instance, assessing the vividness of conscious scenes, or self-consciousness, or Theory of Mind [Vygotsky, 1980] abilities would be incomplete).

- **R6.** The measure should not rely exclusively on behavioral criteria (third person), inner machinery should also be inspected for architecture-based and information processing criteria (this will also prevent conscious-like pre-programed behaviors to fool the measure).

Taking into consideration these requirements we can review existing measures of consciousness and analyze what accounts are closer to meet them all, and why some requirements are not yet fulfilled.

3. Existing Measures of Consciousness and their Application to Machine Consciousness

A detailed review of measures of consciousness is out of the scope of this paper, for a comprehensive review and discussion of measures see [Seth *et al.*, 2008]. In this section we will focus exclusively on the applicability of the most salient measures of consciousness in the domain of MC. We will use the requirements defined above to evaluate the applicability of these measures to machine consciousness implementations.

Clinical diagnosis of disorders of consciousness in humans is usually based on neuro-behavioral criteria [Schnakers *et al.*, 2009]. Related behavioral measures, like the Glasgow Coma Scale [Jennett, 2002] or the more recent JFK Coma Recovery Scale-Revised [Giacino *et al.*, 2004], do not meet requirements R1 and R2 because these measures are specifically designed for humans. Given the limitations of these behavioral scales [Giacino *et al.*, in press], even when applied to humans, neuro-imaging techniques are being appointed as complementary diagnostic tools [Laureys *et al.*, 2004]. However, according to R1, all measures exclusively based on mammalian nervous system, or more specifically, on human brain are not suitable for MC (although they could be of some validity for those MC implementations based on artificial neural systems matching the complexity of the brain). Therefore, all neuro-physiological markers and measures like bispectral index [Rosow and Manberg, 2001], Event-related Cortical Potentials (ERP) [Koivisto and Revonsuo, 2003], neuronal synchrony [Singer and Gray, 1995; Vanderwolf, 2000], etc., are not of direct application to MC. Nevertheless, although these clinical procedures cannot be directly applied to MC, the strategy of combining behavioral assessment methods with inner inspection (like neuro-imaging) can be extrapolated to the field of MC along the lines specified in requirement R6. In fact, behaviors associated with consciousness represent indirect evidence, and it is difficult to differentiate between reflexive and intentional behavior. Therefore, combining behavioral assessment and inner inspection seems to be a good strategy. Discussing specific strategies about inner inspection in MC implementations would be a complete paper on its own, some approaches have been proposed, like looking for software or hardware architectural hallmarks [Sloman, 2002; Arrabales *et al.*, in press], calculating the capability of information

integration of the system [Tononi, 2004; Koch and Tononi, 2008], or looking for the presence of axiomatic properties [Aleksander and Dunmall, 2003].

Given the obvious limitation of clinical diagnosis behavioral scales in their applicability to MC, other behavioral approaches can be explored in order to be combined with inner inspection. One common problem with classical behavioral approaches, like the Turing test [Turing, 1950], is that conditions to pass the test are too strong, and indeed only applicable to human-level consciousness. In other words, the Turing test does not comply with requirement R4 (neither with R5 and R6), not being suitable for measuring different aspects or lower levels of consciousness. As in the Turing test, accurate verbal report is usually applied to assess consciousness in humans. However, this criterion is too strong for machine or animal consciousness. Nevertheless, reportability of mental contents with grounded meaning is a sign of consciousness [Haikonen, 2007], and simpler forms of mental content report could be used in machines. This will imply a redefinition of first person approaches adapted to MC, with the aim to fulfill the requirements specified above. In general, incrementally demanding and content-specific behavioral tests have to be designed in order to fulfill requirements R4 and R5. *ConsScale* is an attempt to meet these requirements, however it is a scale focused on the functionality of cognitive abilities associated with consciousness, and does not provide an account for the phenomenal dimension [Arrabales *et al.*, in press].

In terms of the Information Integration Theory of consciousness [Tononi, 2004], information integration is an indicator of the level of phenomenal consciousness. In relation with this account, the measures of dynamical complexity [Seth, 2009] are not based exclusively in the notion of integration (unity of conscious experience), but in the combination of integration and differentiation (ability to discriminate conscious experiences amongst a vast repertoire of possible scenes). Note that in the context of dynamical complexity the concepts of integration and differentiation refer to the informational value of conscious scenes. While these measures that assess the balance between integration and differentiation provide a characterization of the information complexity in the system, behavioral tests provide an indication of the effective functionality derived from the cognitive capabilities of the subject. As pointed out above, if complexity and functionality are to develop together in MC implementations (although highly complex implementations without useful functionality are possible), a suitable measure should combine these two accounts.

4. Conclusions

In this proposal we have tried to define a practical framework for the problem of measuring consciousness in machines. Although the approach is, of necessity, incomplete, we believe it is practical in terms of applicability and enhancement. A practical and scientifically plausible measure for MC should integrate all the aspects discussed above. Taking just one aspect of consciousness as a canonical reference for the assessment of the level of consciousness of artificial systems would constitute a

partial and biased evaluation. The complexity and multidimensional characterization of consciousness cannot be neglected in the design of a good measure for MC. The design of a comprehensive measure of machine consciousness calls for the integration of first and third person approaches, behavioral and non-behavioral measures, phenomenal and access aspects. Measuring consciousness using a single one-dimensional measure is too reductionist. A good comparative analysis of MC implementations requires R5 to be fulfilled; as each implementation might have different strengths in different aspects of consciousness.

Acknowledgments

We wish to thank Anil K. Seth and Owen Holland for their helpful comments and critique. This work has been supported by the Grant CICYT TRA-2007-67374-C02-02.

References

Aleksander, I. and Dunmall, B. [2003] "Axioms and tests for the presence of minimal consciousness in agents," *Journal of Consciousness Studies* **10**, 7–18.

Arrabales, R., Ledezma, A. and Sanchis, A. [2009] "Establishing a roadmap and metrics for conscious machines development," *Proceedings of the IEEE 8th Conference on Cognitive Informatics*, Hong Kong, pp. 94–101.

Arrabales, R., Ledezma, A. and Sanchis, A. [in press] "ConsScale: A pragmatic scale for measuring the level of consciousness in artificial agents," *Journal of Consciousness Studies.*

Atkinson, A. P., Thomas, M. S. C. and Cleeremans, A. [2000] "Consciousness: Mapping the theoretical landscape," *Trends in Cognitive Sciences*, pp. 372–382.

Baars, B. J. [1988] *A Cognitive Theory of Consciousness* (Cambridge University Press, Cambridge).

Baars, B. J. [2002] "The conscious access hypothesis: Origins and recent evidence," *Trends in Cognitive Sciences*, pp. 47–52.

Block, N. [1995] "On a confusion about a function of consciousness," *Behavioral and Brain Sciences* **18**, 227–287.

Carruthers, P. [2000] *Phenomenal Consciousness: A Naturalistic Theory* (Cambridge University Press, Cambridge).

Chalmers, D. [1995] "Facing up to the problem of consciousness," *Journal of Consciousness Studies* **2**, 200–219.

Dennett, D. [1991] *Consciousness Explained* (Little, Brown and Co., Boston).

Dennett, D. [1996] "Facing backwards on the problem of consciousness," *Journal of Consciousness Studies* **3**, 4–6.

Dennett, D. [1997] *Kinds of Minds: Toward an Understanding of Consciousness* (Basic B, New York).

Edelman, D. B., Baars, B. J. and Seth, A. K. [2005] "Identifying hallmarks of consciousness in non-mammalian species," *Consciousness and Cognition* **14**, 169–187.

Giacino, J. T., Kalmar, K. and Whyte, J. [2004] "The JFK coma recovery scale-revised: Measurement characteristics and diagnostic utility," *Archives of Physical Medicine and Rehabilitation* **85**, 2020–2029.

Giacino, J. T., Schnacker, C., Rodriguez-Moreno, D., Kalmar, K., Schiff, N. and Hirsch, J. [in press] "Behavioral assessment in patients with disorders of consciousness: Gold standard or fool's gold?" in *Progress in Brain Research.*

Haikonen, P. [2007] *Robot Brains. Circuits and Systems for Conscious Machines* (Wiley, UK).

Humphrey, N. [1999] *A History of the Mind: Evolution and the Birth of Consciousness* (Springer, New York).
Jennett, B. [2002] The Glasgow coma scale: History and current practice, *Trauma* **4**, 91−103.
Koivisto, M. and Revonsuo, A. [2003] "An ERP study of change detection, change blindness, and visual awareness," *Psychophysiology* **40**, 423−429.
Koch, K. and Tononi, G. [2008] "Can machines be conscious?" in *IEEE Spectrum Special Report: The Singularity.*
Laureys, S., Owen, A. M. and Schiff, N. D. [2004] "Brain function in coma, vegetative state, and related disorders," *The Lancet Neurology* **3**, 537−546.
Manson, N. [2000] "State consciousness and creature consciousness: A real distinction," *Philosophical Psychology* **13**, 405−410.
Revonsuo, A. and Newman, J. [1999] "Binding and consciousness," *Consciousness and Cognition* **8**, 123−127.
Rosow, C. and Manberg, P. [2001] "Bispectral index monitoring," *Anesthesiology Clinics* **19**, 947−966.
Seth, A., Baars, B. J. and Edelman, D. [2005] "Criteria for consciousness in humans and other mammals," *Consciousness and Cognition* **14**, 119−139.
Seth, A. K. [2007] "The functional utility of consciousness depends on content as well as on state," *Behavioral and Brain Sciences* **30**, 106.
Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M. and Pessoa, L. [2008] "Measuring consciousness: Relating behavioral and neuro-physiological approaches," *Trends in Cognitive Sciences* **12**, 314−321.
Seth, A. [2009] "Explanatory correlates of consciousness: Theoretical and computational challenges," *Cognitive Computation* **1**, 50−63.
Schnakers, C., Vanhaudenhuyse, A., Giacino, J., Ventura, M., Boly, M., Majerus, S., Moonen, G. and Laureys, S. [2009] "Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neuro-behavioral assessment," *BMC Neurology* **9**, 35.
Singer, W. and Gray, C. M. [1995] "Visual feature integration and the temporal correlation hypothesis," *Annual Review of Neuroscience* **18**, 555−586.
Sloman, A. [2002] "Architecture-based conceptions of mind," *In the Scope of Logic, Methodology, and Philosophy of Science* **2**, 403−427.
Tononi, G. [2004] "An information integration theory of consciousness," *BMC Neuroscience* **5**.
Turing, A. [1950] "Computing Machinery and Intelligence," in *Mind.*
Vanderwolf, C. H. [2000] "Are neocortical gamma waves related to consciousness?" *Brain Research* **855**, 217−224.
Vygotsky, L. S. [1980] *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, Harvard).