THE COGNITIVE DEVELOPMENT OF MACHINE CONSCIOUSNESS IMPLEMENTATIONS

RAÚL ARRABALES*, AGAPITO LEDEZMA and ARACELI SANCHIS

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

The progress in the machine consciousness research field has to be assessed in terms of the features demonstrated by the new models and implementations currently being designed. In this paper, we focus on the functional aspects of consciousness and propose the application of a revision of *ConsScale* a biologically inspired scale for measuring cognitive development in artificial agents in order to assess the cognitive capabilities of machine consciousness implementations. We argue that the progress in the implementation of consciousness in artificial agents can be assessed by looking at how key cognitive abilities associated to consciousness are integrated within artificial systems. Specifically, we characterize *ConsScale* as a partially ordered set and propose a particular dependency hierarchy for cognitive skills. Associated to that hierarchy a graphical representation of the cognitive profile of an artificial agent is presented as a helpful analytic tool. The proposed evaluation schema is discussed and applied to a number of significant machine consciousness models and implementations. Finally, the possibility of generating qualia and phenomenological states in machines is discussed in the context of the proposed analysis.

Keywords: Cognitive development; machine consciousness; ConsScale; measures of consciousness.

1. Introduction

Artificial systems created as part of current machine consciousness research efforts are usually inspired by certain aspects of biological organisms. However, the specific inspiring models and the particular way in which they are implemented may differ greatly from one system to another. Consequently, it is not straightforward to characterize the cognitive capabilities of an artificial architecture in such a way that it can be put in a general context, i.e., compared with other implementations based on different principles. The root of the problem lies in the fact that different perspectives and aspects are usually confusedly merged under the concept of consciousness [Block, 1995].

In this work, we focus on the problem of identifying the most important cognitive functions associated with consciousness and the question of how these functions can

be effectively integrated in order to build a human like agent. The definition of a generic framework for the evaluation and characterization of the cognitive development of an artificial agent can be beneficial not only for the comparative analysis of existing models, but also for the planning of a roadmap for future implementations [Arrabales *et al.*, 2009]. *ConsScale* is a proposal intended to define such a framework using architectural and behavioral criteria [Arrabales *et al.*, 2010]. While most of the existing consciousness metrics proposals are based on low level information integration measures [Tononi, 2004; 2008; Seth, 2005], *ConsScale* is in contrast based on higher level functional aspects of the system. It is important to remark that we do not disregard information integration as a key property of conscious systems; in fact, we aim to characterize how effective information integration and inter function synergies can contribute to the generation of conscious like behaviors. In short, while measures like Φ look exclusively at the information integration capabilities of the system [Tononi, 2008], *ConsScale* aims at specifying — at the functional level — how well this integration translates into adaptive behavior. As argued elsewhere [Arrabales *et al.*, 2009], both information integration and behavioral measures should be combined in order to provide a comprehensive evaluation method for potentially conscious machines.

The main conceptual tool we use for the characterization of the cognitive development of artificial creatures is the definition of a partially ordered set of cognitive skills. This taxonomy — based on the development of consciousness — is used to analyze, classify, and compare the cognitive profile of both unimplemented computational models of consciousness and extant machine consciousness implementations.

In the following we briefly describe the levels of consciousness defined in *ConsScale* and discuss a revised cognitive hierarchy based on dependency relations (Sec. 2); then, we describe the main tools associated with the scale and describe the associated rating methodologies (Sec. 3). After that, we introduce the new proposal for graphical cognitive profiling, using it to analyze some salient machine consciousness models and implementations (Sec. 4). Finally, we draw some conclusions on the former analysis and discuss the implications in terms of the generation of qualia and phenomenal consciousness assessment for artificial agents (Sec. 5).

2. Levels of Consciousness

ConsScale levels are defined using both architectural and functional criteria. In this paper, we will focus mainly on the cognitive (functional) capabilities for the discussion on the assessment of the global level of cognitive development of an artificial agent. Although a total of 13 levels are defined in *ConsScale* (from level −1 to level 11, including level 0), only the most common 10 levels are considered here: 2 — *Reactive*, 3 — *Adaptive*, 4 — *Attentional*, 5 — *Executive*, 6 — *Emotional*, 7 — *Self Conscious*, 8 — *Empathic*, 9 — *Social*, 10 — *Human Like*, and 11 — *Super Conscious*. Table 1 summarizes the cognitive skills required in these levels. Each level defines a set of generic cognitive skills ($CS_{i,j}$) that must be satisfied. Note that agents can only qualify

Table 1. *ConsScale* levels 2 to 11.

Level (L_i)	Cognitive skills
2	$CS_{2,1}$: Fixed reactive responses ("reflexes").
3	$CS_{3,1}$: Autonomous acquisition of new adaptive reactive responses. $CS_{3,2}$: Usage of proprioceptive sensing for embodied adaptive responses. $CS_{3,3-5}$ [a]: Selection of relevant sensory/motor/memory information. $CS_{3,6}$ [a]: Evaluation (positive or negative) of selected objects or events. $CS_{3,7}$ [a]: Selection of what needs to be stored in memory.
4	$CS_{4,1}$: Trial and error learning. Re-evaluation of selected objects or events. $CS_{4,2}$: Directed behavior toward specific targets like following or escape. $CS_{4,3}$: Evaluation of the performance in the achievement of a single goal. $CS_{4,4}$: Basic planning capability: calculation of next n sequential actions. $CS_{4,5}$: Ability to build depictive [Aleksander and Dunmall, 2003] representations of percepts for each available sensory modality.
5	$CS_{5,1}$: Ability to move back and forth between multiple tasks. $CS_{5,2}$: Seeking of multiple goals. $CS_{5,3}$: Evaluation of the performance in the achievement of multiple goals. $CS_{5,4}$: Autonomous reinforcement learning (emotional learning). $CS_{5,5}$: Advanced planning capability considering all active goals. $CS_{5,6}$ [b]: Ability to generate selected mental content with grounded meaning [Haikonen, 2007] integrating different modalities into differentiated explicit percepts [Tononi, 2008].
6	$CS_{6,1}$: Self-status assessment (background emotions). $CS_{6,2}$: Background emotions cause effects in agent's body. $CS_{6,3}$: Representation of the effect of emotions in organism and planning (feelings) [Damasio, 1999]. $CS_{6,4}$: Ability to hold a precise and updated map of body schema. $CS_{6,5}$: Abstract learning (learned lessons generalization). $CS_{6,6}$ [b]: Ability to represent a flow of integrated percepts including self-status.
7	$CS_{7,1-3}$: Representation of the relation between self and perception/action/feelings. $CS_{7,4}$: Self-recognition capability. $CS_{7,5}$: Advance planning including the self as an actor in the plans. $CS_{7,6}$: Use of *imaginational* states in planning. $CS_{7,7}$: Learning of tool usage. $CS_{7,8}$ [b]: Ability to represent and self-report mental content (continuous inner flow of percepts inner imagery).
8	$CS_{8,1}$: Ability to model others as subjective selves. $CS_{8,2}$: Learning by imitation of a counterpart. $CS_{8,3}$: Ability to collaborate with others in the pursuit of a common goal. $CS_{8,4}$: Social planning (planning with socially aware plans). $CS_{8,5}$: Ability to make new tools. $CS_{8,6}$ [b]: Inner imagery is enriched with mental content related to the model of others and the relation between the self and other selves.
9	$CS_{9,1}$: Ability to develop Machiavellian strategies like lying and cunning. $CS_{9,2}$: Social learning (learning of new Machiavellian strategies). $CS_{9,3}$: Advanced communication skills (accurate report of mental content as basic inner speech). $CS_{9,4}$: Groups are able to develop a culture. $CS_{9,5}$ [a]: Ability to modify and adapt the environment to agent's needs.

Table 1. (*Continued*)

Level (L_i)	Cognitive skills
10	$CS_{10,1}$: Accurate verbal report. Advanced linguistic capabilities. Human-like inner speech. $CS_{10,2}$: Ability to pass the Turing test. $CS_{10,3}$: Groups are able to develop a civilization and advance culture and technology.
11	$CS_{11,1}$: Ability to manage several streams of consciousness.

[a]This *CS* has been changed in this revised version of *ConsScale*.
[b]This *CS* has been added in this revised version of *ConsScale*.

as a given level n *if and only if* all lower levels are also fully satisfied. In order to apply the scale to a real world problem, these *CS* need to be grounded (or instantiated) to actual behavioral tests, which could be evaluated via third person observations (see [Arrabales *et al.*, 2009] for a *ConsScale* instantiation in the domain of first person shooter game synthetic characters or "bots").

In the revised version of *ConsScale* presented here, the relations between the different *CS* have been formalized considering a finite partially ordered set (poset) [Stanley, 2000], and can be visualized through its Hasse diagram (Fig. 1). The *CS* hierarchy is based on a strict partial order binary relation "$<$" that represents "cognitive dependency". Therefore, the set of all *CS* in *ConsScale* (CCS) partially ordered by the relation cognitive dependency can be regarded as a poset (CCS, $<$). For instance, $CS_{6,4} < CS_{7,4}$ (represented in Fig. 1 by an upward arrow from vertex $CS_{6,4}$ to vertex $CS_{7,4}$) means that $CS_{7,4}$ covers $CS_{6,4}$. In other words, *self recognition capability* ($CS_{7,4}$) requires the *ability to hold a precise and updated map of body schema* ($CS_{6,4}$). Analogously, other dependency relations have been identified between the rest of skills as illustrated in Fig. 1. The detailed explanation of each dependency relation cannot be included in this paper due to space limitations and will need to be addressed elsewhere.[1] As a general rule, current *CS* definition and associated hierarchy satisfies that no higher level skill is required to attain a lower level skill.

CCS is not a totally ordered set because not all skills are comparable. In fact, *ConsScale* levels are defined as subsets of incomparable skills. The dependency relations have been established considering human ontogeny and biological phylo geny [Arrabales *et al.*, 2010].

The poset $(\mathrm{CCS}, <)$ is composed of a number of inter related subsets that rep resent the development and composition of specific cognitive functions. If we con sider, for instance, Theory of Mind (ToM) [Lewis, 2003] the following partial order is included in CCS:

$$CS_{6,1\ \ 6}(\text{I know}) < CS_{7,1\ \ 5}(\text{I know I know}) < CS_{8,1\ \ 4}(\text{I know you know}) < CS_{9,1\ \ 2}(\text{I know you know I know})$$

[1]See http://www.consscale.com for supporting data and additional information about the scale.

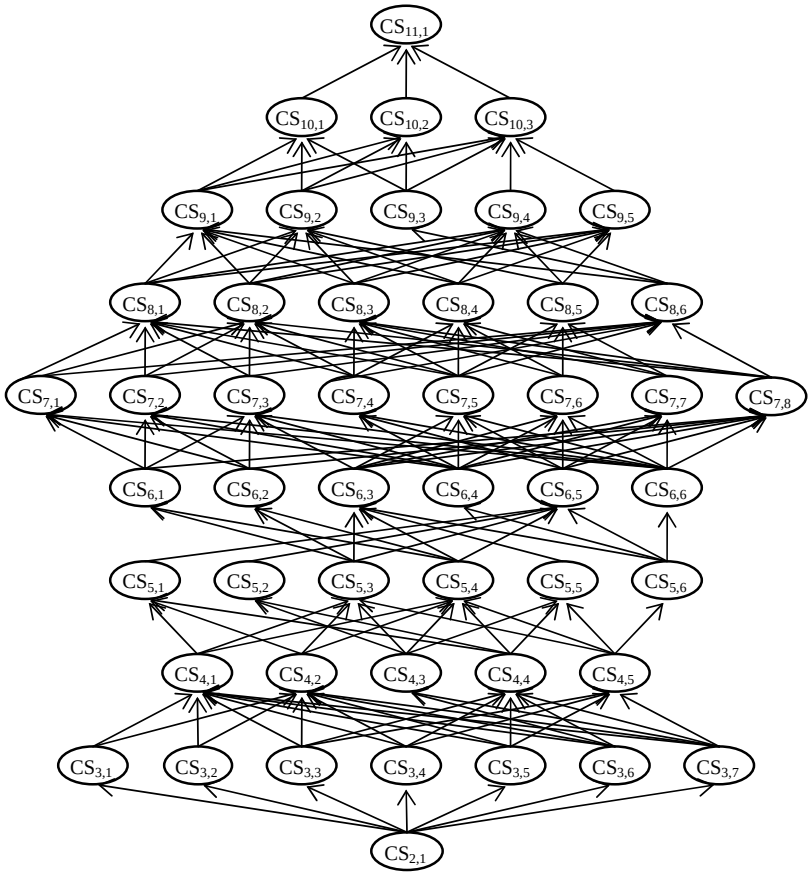


Fig. 1. Hasse diagram of the poset (CCS, $<$) that represents dependency relations between cognitive skills.

The same principle applies for other cognitive functions like executive function, the modulating function of emotions, and generation of inner speech and accurate verbal report [Arrabales *et al.*, 2010].

3. Evaluating Artificial Agents Using *ConsScale*

As we have argued, a machine consciousness implementation can be studied and evaluated with the aim to find out which *CS* from the former list (see Table 1) are present. However, a comprehensive characterization of the degree of cognitive development of the implementation calls for the combination of the results of all levels. In other words, an integrative measure is required.

Two different cognitive characterization tools are described in the following. The first one consists on the application of a quantitative score and has been already discussed in detail elsewhere [Arrabales *et al.*, 2009]. The second one is a proposal intended to enhance the cognitive power characterization that *ConsScale* can offer, and is based on graphical cognitive profile representations.

3.1. *ConsScale quantitative score*

The *ConsScale* Quantitative Score (CQS) is an assessment tool associated with the scale. It is intended to provide a numerical value as an indication of the cognitive power of the implementation being evaluated. The CQS is calculated in three steps:

(i) L_i, or level i compliance, provides a measure (from 0.0 to 1.0) which follows an exponential curve as a means to represent the synergy between different skills within the same level, i.e., the greater the number of *CS* fulfilled, the greater will be the contribution of additional skills to the overall behavior of the agent.

(ii) CLS, or Cumulative Level Score, combines all L_i into one single aggregated value (from 0.0 to $\sim$1.55). This score follows a logarithmic progression which prevents the final score to be distorted by the combined effect of large L_i scores in higher levels and poor L_i scores in lower levels (e.g., implementations good at levels 5 and 6 but showing poor results in lower levels should not be awarded high scores).

(iii) CQS provides a single value (from 0.0 to 1000) that indicates the cumulative synergy produced by the integration of cognitive skills across all levels. CQS is designed as an exponential curve priming those implementations which follow the developmental path implicitly defined in *ConsScale* level ordering (see Fig. 2).

The mathematical procedure and details about the calculation of CQS can be found in [Arrabales *et al.*, 2009]. Additionally, a CQS calculator is available online at the *ConsScale* website.

3.2. *ConsScale rating approaches*

ConsScale is based on the hypothesis that effective integration of the cognitive abilities listed in Table 1 (and associated architectural components) is required in

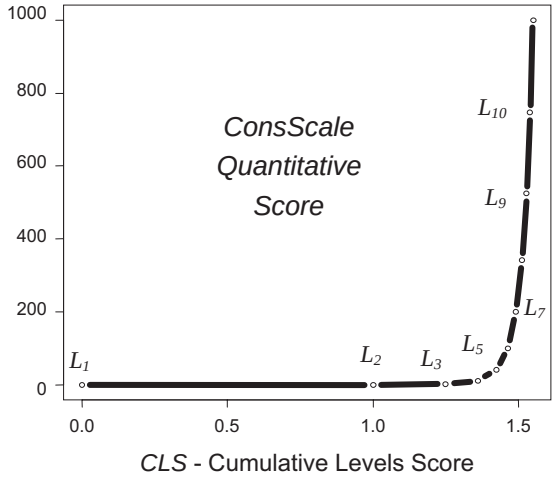


Fig. 2. Possible CQS values as a function of CLS.

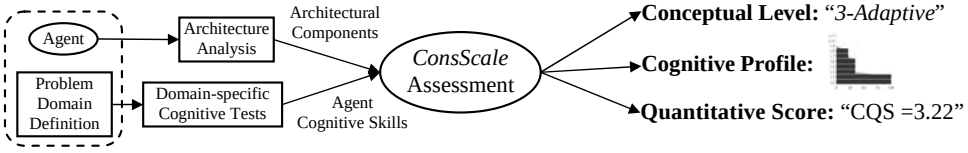


Fig. 3. *ConsScale* Standard Evaluation Process (SEP).

order to develop behaviors associated with conscious beings. In order to assess the overall cognitive development of an agent, two approaches can be applied:

- The *ConsScale* Standard Evaluation Process (SEP) is oriented to existing implemented agents and provides an accurate and compelling measure of the level of cognitive development (see Fig. 3).
- The *ConsScale* Simplified Rating Process (SRP) provides a quick approximation of the potential level of cognitive development of either an existing agent or even a computational model not yet implemented (see Fig. 4).

Performing a SEP requires the actual agent and a particular problem domain definition for testing. As mentioned above, rating is based on architectural com ponents and cognitive skills. Architectural components of the agent are identified through internal inspection of the implementation. Cognitive skills present in the agent are assessed thanks to the definition and execution of specific cognitive tests adapted to the established problem domain. Once the list of architectural com ponents and cognitive skills has been determined for the particular agent, the *ConsScale* metrics can be applied in order to obtain the nominal level of functional consciousness, the cognitive graphical profile, and the CQS score.

Note that comprehensive cognitive tests have to be devised for each cognitive skill. These tests have to be designed in such a way that they validate the integrative and developmental inspiration of the scale. In other words, higher level cognitive tests will require the presence and effective integration of all of lower cognitive abilities (according to the "<" relation defined) in order to be passed. See [Arrabales *et al.*, 2009] for an example of SEP in the domain of first person shooter computer game bots.

The SRP assumes the presence of architectural components and cognitive skills just by looking at the design blueprints of the system. Therefore, there is no need to perform any test or to use any domain specific instantiation of the scale. Of course, the rating obtained following this procedure is not accurate and can be considered as

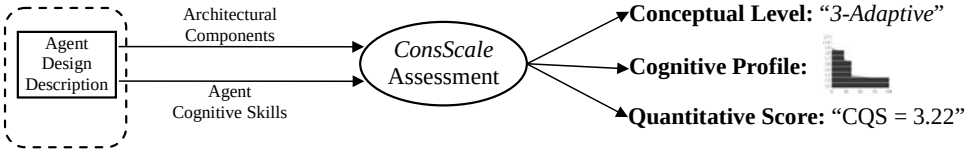


Fig. 4. *ConsScale* Simplified Rating Process (SRP).

just a vast approximation (and probably too optimistic). However, SRP could be also used as a way to assess the potential of a machine consciousness model early during its initial design phase.

In the case of implemented agents, SEP should be performed in order to obtain an accurate and realistic measure. Nevertheless, the SRP provides a conceptual tool to evaluate the potential *ConsScale* level of a cognitive architecture even at design time, before any implementation of the model exists (see Table 2 for a comparison between SEP and SRP).

4. Machine Consciousness Implementation Evaluation Examples

Although having a single quantitative measure like CQS is useful for a quick charac terization and evaluation, it lacks rich representation capabilities. For that reason, we have proposed the complementary use of graphical representations of cognitive profiles [Arrabales *et al.*, 2009]. In order to represent the cognitive profile of an agent in terms of *ConsScale* the particular L_i scores have to be considered. Note that both CLS and CQS are one dimensional parameters, calculated as a function of the multidimensional L_i; therefore, $L_i (i \in \{2-11\})$ are the parameters to be used for a graphical representation that preserves the multidimensional richness of *ConsScale* levels definition (for the sake of clarity, *ConsScale* levels $-1, 0$, and 1 have been excluded).

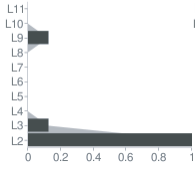
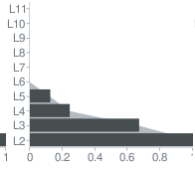
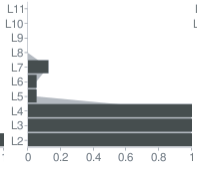
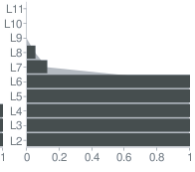
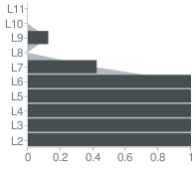
Although in former work we have used radar charts [Arrabales *et al.*, 2009], in this paper we have decided to use horizontal bar charts as a compact and meaningful layout for the representation of the L_i values. The hierarchical nature of the scale is well represented using this arrangement, where lower levels are placed in the bottom and higher levels on top. Each bar represents the degree of accomplishment in the corresponding *ConsScale* level. Table 3 illustrates the use of the graphical cognitive profiles applied to the comparative analysis of several systems. In order to provide an approximate but illustrative view of current state of the art, the following machine consciousness models or implementations have been analyzed using the SRP:

- Eliza: one of the first chatterbots [Weizenbaum, 1966].

Table 2. Comparison between *ConsScale* simplified and standard rating processes.

	Standard Rating (SEP)	Simplified Rating (SRP)
Applicability	Only extant implementations.	Models, designs, implementations.
Accuracy	High (realistic metric).	Low (potential, optimistic metric).
Cost	High (internal inspection of the implementation, cognitive test design and execution).	Low (features are inferred directly).
Problem Domain	Domain-dependent.	Domain-independent.
Required Resources	Suitable testing environment, test procedures and tools, data collect ion and inspection tools.	Detailed description of the system.
Output	*ConsScale* level, cognitive profile, CQS.	*ConsScale* level, cognitive profile, CQS.

Table 3. Summary of SRP results. From top to bottom: name of system, satisfied cognitive skills, overall quantitative score, graphical cognitive profile, and *ConsScale* conceptual level. ELIZA is basically a reactive agent designed to detect and select keywords in the input and, using a script and pattern matching technique, provide a response in the form of accurate verbal report ($CS_{9,3}$). Although this agent presents one of the highest level of cognitive skills, the final CQS is low because *ConsScale* primes a developmental integration of cognitive abilities. In this particular case, it does not matter how good the agent is at producing well-formed linguistic reports: if the "mental" content reported is not created by a suitable combination of lower-level cognitive abilities, the scale cannot consider the agent as cognitively advanced. UT Bot complied with some features of levels 3, 4, and 5, however it is rated as level 2 because *ConsScale* requires the complete fulfillment of lower levels in order to qualify as a given level i. The CQS for a pure reactive agent is 0.18. However, the UT Bot score (0.51) indicates that some additional cognitive features are in place (as can be noticed in its associated cognitive profile). Nevertheless, UT Bot is far from a level 4 agent who would score 12.21 or more. The CRONOS minimal architecture for functional imagination (MAFI) is rated as level 4. However, being a minimal architecture implementation, the proposal is promising in terms of achieving higher scores. Actually, the multiple step architecture with memory [Marques, 2009] enhance this model including $CS_{5,1}$. Both LIDA and Haikonen's architecture are roughly equivalent in terms of *ConsScale*. Nevertheless, a comprehensive testing of full implementations would be required in order to see if such machine consciousness implementations could be promoted to *ConsScale* level 7 (*self-conscious*).

ELIZA	**CERA UT Bot**	**CRONOS MAFI**	**LIDA**	**Haikonen**
$CS_{2\,1}$; $CS_{3\,3}$; $CS_{3\,4}$; $CS_{3\,5}$; $CS_{9,3}$.	$CS_{2\,1}$; $CS_{3\,1}$; $CS_{3\,2}$; $CS_{3\,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{4,1}$; $CS_{4,5}$; $CS_{5,2}$; $CS_{5,4}$.	$CS_{2\,1}$; $CS_{3\,1}$; $CS_{3\,2}$; $CS_{3\,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{3,7}$; $CS_{4\,1}$; $CS_{4\,2}$; $CS_{4\,3}$; $CS_{4\,4}$; $CS_{4,5}$; $CS_{5,4}$; $CS_{6,4}$; $CS_{7,1}$; $CS_{7,2}$; $CS_{7,5}$; $CS_{7,6}$.	$CS_{2\,1}$; $CS_{3\,1}$; $CS_{3\,2}$; $CS_{3\,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{3,7}$; $CS_{4\,1}$; $CS_{4\,2}$; $CS_{4\,3}$; $CS_{4\,4}$; $CS_{4,5}$; $CS_{5,1}$; $CS_{5,2}$; $CS_{5,3}$; $CS_{5\,4}$; $CS_{5\,5}$; $CS_{5\,6}$; $CS_{6\,1}$; $CS_{6,2}$; $CS_{6,3}$; $CS_{6,4}$; $CS_{6,5}$; $CS_{6\,6}$; $CS_{7\,1}$; $CS_{7\,2}$; $CS_{7\,3}$; $CS_{7,6}$; $CS_{8,1}$.	$CS_{2\,1}$; $CS_{3\,1}$; $CS_{3\,2}$; $CS_{3\,3}$; $CS_{3,4}$; $CS_{3,5}$; $CS_{3,6}$; $CS_{3,7}$; $CS_{4\,1}$; $CS_{4\,2}$; $CS_{4\,3}$; $CS_{4\,4}$; $CS_{4,5}$; $CS_{5,1}$; $CS_{5,2}$; $CS_{5,3}$; $CS_{5\,4}$; $CS_{5\,5}$; $CS_{5\,6}$; $CS_{6\,1}$; $CS_{6,2}$; $CS_{6,3}$; $CS_{6,4}$; $CS_{6,5}$; $CS_{6\,6}$; $CS_{7\,1}$; $CS_{7\,2}$; $CS_{7\,3}$; $CS_{7,4}$; $CS_{7,8}$; $CS_{9,3}$.
CQS: **0.19**	CQS: **0.51**	CQS: **12.37**	CQS: **102.27**	CQS: **114.39**
				
2 (*reactive*)	2 (*reactive*)	4 (*attentional*)	6 (*emotional*)	6 (*emotional*)

- UT2004 Adaptive Bot: an Unreal Tournament 2004 autonomous bot implemented using the CERA CRANIUM cognitive architecture [Arrabales *et al.*, 2009].
- Functional Imagination on CRONOS/SIMNOS: implementation of a functional imagination mechanism that allows an embodied agent to simulate its own actions and their sensory consequences internally, and to extract behavioral benefits from doing so [Marques and Holland, 2009].
- LIDA model: LIDA is a (not yet fully implemented) comprehensive computational model of cognition primarily based on the Global Workspace Theory [Franklin *et al.*, 2007; Baars and Franklin, 2009].
- Haikonen's Cognitive Architecture: cognitive architecture based on distributed signal representations and Haikonen Associative Neurons [Haikonen, 2007].

Table 3 summarizes the preliminary evaluation results after applying SRP. Note that simplified rating provides just an approximation of what could be the real *ConsScale* level of an implementation. The rating obtained for models which have not yet been fully implemented will have to be confirmed in the future by the application of the SEP to the corresponding implementations. For implementations or models which consider a developmental period, the rating considers the potential final *ConsScale* level that they would achieve at the end of their developmental period. See Table 3 for a comparative analysis of the five machine consciousness systems being discussed.

Looking at the cognitive profiles in Table 3 it can be easily noticed that all the analyzed machine consciousness models essentially follow the developmental path outlined by *ConsScale* hierarchical levels. This is indeed the expected result due to the existing dependencies between the skills arranged at different levels. Nevertheless, machine consciousness models (as well as biological organisms) might exist that present "atypical" cognitive profiles, e.g., an autistic person or an artificial agent specifically pre programed to recognize its own specular image (without fulfilling lower level skills). Usually, these atypical cognitive profiles appear in nature due to brain injury or genetic diseases. However, in the case of artificial systems it might indicate either a task oriented design or even the presence of pre programed behaviors conceived to fool classical cognitive tests. *ConsScale* CQS represents the cognitive hierarchical dependency and applies a synergistic weighting function in order to account for such systems providing a fair measure of their overall cognitive power.

5. Conclusions

The analysis of the selected machine consciousness models indicates that *ConsScale* profiles associated to the corresponding implementations — after applying the SEP — would have good scores only in the lower section of the chart. Looking at the preliminary results obtained using the SRP, the following conclusions can be drawn:

- Although the detailed *CS* dependency relations between adjacent levels (illus trated in Fig. 1) can be a subject of controversy and might require further refinement, it is clear that functions located in highest levels do require the effective realization and integration of lowest level functions. Hence, at least from a coarse grained perspective, the cognitive hierarchy proposed in *ConsScale* is supported by the engineering constraints found in machine consciousness implementations (as well as the equivalent dependencies observed in biological phylogeny).
- Similarly, the analysis of the selected systems confirms that higher level skills are not required to attain lower level skills, thus supporting the upward ordering relations defined in *ConsScale*.
- Although a lot of work still needs to be done in order to build real implementations able to successfully cope with *ConsScale* lower levels, the actual challenge in the field of machine consciousness is to create new artificial creatures whose cognitive

profiles tend to fill the upper half of the chart (while keeping high scores in the lower half).

As shown in this paper, the proposed evaluation methods (SEP and SRP) are valid for very dissimilar implementations, thus allowing comparative analysis across all possible models that might arise in the domain of machine consciousness. However, these methods have some drawbacks: while SEP permits an accurate analysis of a given implementation, it is of necessity domain dependent, therefore an accurate comparative analysis can only be performed between systems designed to work in the same context. In order to compare systems intended to be used in different domains — as in the case of this paper — the SRP has to be applied. Regrettably, this method provides just an approximated evaluation, which might be too sensible to *CS* arbitrary interpretations in the context of each particular system. For instance, an agent is said to comply with $CS_{3,4}$ if it is able to "*adaptively select relevant motor information*". This could mean different things in different contexts, and involve much more engineering effort in some domains than in others. For the agent UT Bot $CS_{3,4}$ it is translated into "*the ability of the bot to discard actions that are not suitable for the current situation*", like firing against walls while running away from an enemy [Arrabales *et al.*, 2009]. For the functional imagination architecture, $CS_{3,4}$ could be translated into the "*ability to pre select motor actions directed towards the goal*" [Marques and Holland, 2009], like moving the arms in the direction of the object that has to be knocked down. Whereas implementing and testing these two different behaviors might imply quite different designs and techniques, their cognitive significance is equivalent from the point of view of *ConsScale*. In other words, *ConsScale* SRP does not take into account the complexity of the application domain; therefore the metrics obtained in this work are not sensitive to robustness versus brittleness in agents. As mentioned above, the *ConsScale* SEP has to be used (instead of SRP) in order to obtain a fair and accurate comparative metric — at the cost of constraining the evaluation to a specific problem domain.

Another problem is related to the particular evaluation of each *CS*. While the fulfillment of a given skill is now considered as a binary property, real implementations generally present a blurred boundary between behaviors that could be considered as satisfying or not certain *CS*. For instance, in the case of $CS_{7,4}$, the mirror test could be used to evaluate the agent. A typical outcome of the test could be that the agent is able to pass the mirror test with an accuracy of 70% [Takeno *et al.*, 2005]. Arbitrarily translating this sort of results into a binary property obviously induces noise and ambiguity in the metric. This effect could be diminished by considering partial fulfillment of *CS* and/or fuzzy logic in the calculation of L_i parameters.

Although the proposed scale does not explicitly address the problem of phenomenal consciousness assessment, it could be argued that some correlation might exist between the assessed functional synergy and the probability of having phenomenological states. While the functional synergy might not be required for the generation

of phenomenological states, it seems to be a requirement for the formation of qualia, the integrated content of subjective experience.

As pointed out by Haikonen [Haikonen, 2009], qualia is the way in which sensory information manifest itself in mind, therefore the production of "artificial qualia" in machines has to be considered when assessing the degree of consciousness of a machine. In this regard, we are currently investigating the correlations between cognitive processes defined in *ConsScale* and the generation and development of qualia. Specifically, the partial order: $CS_{4,5} < CS_{5,6} < CS_{6,6} < CS_{7,8} < CS_{8,6} < CS_{9,3} < CS_{10,1}$. Taking into account this CCS partial order and the models being analyzed we have found that current machine consciousness designs are also following the *Cons Scale* path for the creation of artificial qualia. For instance, in the case of the LIDA model, the contents of the conscious broadcast are said to constitute the artificial qualia of the agent. In Haikonen's architecture, the mechanism for direct and trans parent perception is considered essential for the potential creation of artificial qualia.

Acknowledgments

We wish to thank the anonymous reviewers for their valuable suggestions. We are also grateful to Pentti Haikonen, Stan Franklin, Hugo Gravato and Owen Holland for the useful feedback they provided about their systems. 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 8th IEEE International Conference on Cognitive Informatics*, Hong Kong, pp. 94 101.

Arrabales, R., Ledezma, A. and Sanchis, A. [2009] "Strategies for measuring machine consciousness," *International Journal of Machine Consciousness* **1**, 193 201.

Arrabales, R., Ledezma, A. and Sanchis, A. [2009] "Assessing and characterizing the cognitive power of machine consciousness implementations," *AAAI Fall Symposium Series.*

Arrabales, R., Ledezma, A. and Sanchis, A. [2009] "Towards conscious like behavior in com puter game characters," *IEEE Symposium on Computational Intelligence and Games*, Milano, Italy.

Arrabales, R., Ledezma, A. and Sanchis, A. [2010] "ConsScale: A pragmatic scale for measuring the level of consciousness in artificial agents," *Journal of Consciousness Studies* **17**, 131 164.

Baars, B. J. and Franklin, S. [2009] "Consciousness is computational: The LIDA model of global workspace theory," *International Journal of Machine Consciousness* **1**, 23 32.

Block, N. [1995] "On a confusion about a function of consciousness," *Behavioral Brain Science*, pp. 227 287.

Damasio, A. R. [1999] *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Heinemann, London).

Franklin, S. *et al.* [2007] "LIDA: A computational model of global workspace theory and developmental learning," *AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches.*

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

Haikonen, P. O. A. [2009] "Qualia and conscious machines," *International Journal of Machine Consciousness* **1**, 225 234.

Lewis, M. [2003] "The emergence of consciousness and its role in human development," *Annual NY Academy of Science* **1001**, 104 133.

Marques, H. G. [2009] "Architectures for embodied imagination," Ph.D. Thesis, University of Essex.

Marques, H. G. and Holland, O. [2009] "Architectures for functional imagination," *Neuro computing* **72**, 743 759.

Seth, A. K. [2005] "Causal connectivity of evolved neural networks during behavior," *Network: Computation in Neural Systems* **16**, 35.

Stanley, R. P. [2000] *Enumerative Combinatorics* (Cambridge University Press, Cambridge, UK).

Takeno, J., Inaba, K. and Suzuki, T. [2005] "Experiments and examination of mirror image cognition using a small robot," *CIRA 2005*, pp. 493 498.

Tononi, G. [2004] "An information integration theory of consciousness," *BMC Neuroscience* **5**, 42.

Tononi, G. [2008] "Consciousness as integrated information: A provisional manifesto," *Biological Bulletin* **215**, 216 242.

Weizenbaum, J. [1966] "ELIZA a computer program for the study of natural language com munication between man and machine," *Communication of ACM* **9**, 36 45.