Consciousness, Meaning and the Future Phenomenology

Ricardo Sanz[1,2], Carlos Hernández[1] and Guadalupe Sánchez[1]

Abstract. Phenomenological states are generally considered sources of intrinsic motivation for autonomous biological agents. In this paper we will address the issue of exploiting these states for robust goal-directed systems. We will provide an analysis of consciousness in terms of a precise definition of how an agent "understands" the informational flows entering the agent. This model of consciousness and understanding is based in the analysis and evaluation of phenomenological states along potential trajectories in the phase space of the agents. This implies that a possible strategy to follow in order to build autonomous but useful systems is to embed them with the particular, ad-hoc phenomenology that captures the requirements that define the system usefulness from a requirements-strict engineering viewpoint.

1 INTRODUCTION

Research into machine consciousness is justified in terms of the potential increase of functionality [25] but also as a source of experimentation with models of human consciousness to evaluate their value [19].

Even when there are old arguments against the possibility of machine consciousness[3], several attempts at realisations of machine consciousness have been done recently [19]. In some cases, these systems propose a concrete theory of consciousness explicitly addressing artificial agents [15, 10] but in other cases the implementations follow psychological or neural theories of human consciousness developed without considering machines as potential targets for them. This is true, for example in the case of the many implementations of Baars' Global Workspace Theory of consciousness [3, 21, 13, 26].

These are very valuable efforts that help clarify the many issues surrounding consciousness and foster a movement towards making more precise the sometimes too-philosophical terms used in this domain. All these different implementations —if accepted as conscious— may be considered as exemplars in an attempt towards an ostensive definition of *consciousness* that includes humans and maybe also some animals [4].

However as pointed out by Sloman [28] *"pointing at several examples may help to eliminate some misunderstandings by ruling out concepts that apply only to a subset of the examples, but still does not identify a concept uniquely since any set of objects will have more than one thing in common."* In a sense, the only possibility of real, sound advance in machine consciousness is to propose and risk a background theory against to which experiments are done and evidence thrown. This is indeed the path followed by the works previously mentioned of Chella, Haikonen, Franklin, Arrabales or Shanahan. However, most of the approaches are focused on just one aspect of consciousness [5]. The multifarious character of consciousness is an obvious problem.

Indeed, Sloman [28] suggests that the main difficulty that we confront in the research on consciousness and machine consciousness is related to the *polymorphic* nature of the *consciousness* concept. This may seem to imply that trying to tackle several aspects of consciousness -access consciousness, phenomenal consciousness, self-awareness, etc. — in one single shot —a single model, a single robot— is hopeless. This program of addressing consciousness as a whole is also hampered by the semantical flaws that some of the conceptions of consciousness suffer when abstracted from specific contexts.

However, Sloman also recognises that *"perhaps one day, after the richness of the phenomena has been adequately documented, it will prove possible to model the totality in a single working system with multiple interacting components."* This is, boldly, what we try to do inside our long term ASys research program. In order to progress in the systematic engineering of autonomous, robust agents, we will try to make them conscious. And will try to do so by using a *single, general* and *unified* theory of consciousness[4].

The approach taken in this effort directly attacks the polymorphic nature of the concept. We will express general consciousness mechanisms in the form of architectural patterns that will be instantiated in the several forms that are necessary for the specific uses of a particular agent. This approach breaks up the unicity/variety problem of consciousness, leveraging a single structure for different uses.

2 THE REASONS FOR ACTING

The quest for control architectures for artificial autonomous agents confronts a problem concerning the relations between the goals of the agent and the goals of the owner. This is very much connected with the value systems of humans and how these drive their behaviour [23].

Phenomenological states are generally considered sources of intrinsic motivation for autonomous biological agents. At the end of the day, what counts is the phenomenology. What is relevant for the agent is how the internal changes concerning its perception of the world and of itself impacts its experiential state [9].

To be more precise, for us humans, what counts is the integral, i.e. an accumulated value, of the phenomenological states along the lived trajectories —past, present and future. This

[1] Autonomous Systems Laboratory, Universidad Politécnica de Madrid, José Gutierrez Abascal 2, 28006 Madrid, Spain. www.aslab.upm.es. Email: {Ricardo.Sanz, Carlos.Hernandez, Guadalupe.Sanchez}@aslab.org.

[2] Sackler Center for Consciousness Science, University of Sussex, Falmer, East Sussex, UK. www.sussex.ac.uk/sackler.

[3] Paul Ziff, in 1959 said: "*Ex hypothesi* robots are mechanisms, not organisms, not living creatures. There could be a broken-down robot but not a dead one. Only living creatures can literally have feelings." [32]

[4] *Single*, because we are going to propose only one; *general* because we intend it to be of applicability to any kind of system, whether natural or artificial; and *unified* because it shall address all the conceptual spectrum of consciousness (except bogus terms).

is the very foundation for acting —the reasons to act— and the very grounding of ethics. We just care about feeling well and having the right experiences. This may sound a bit selfish but even altruistic behaviour shall be gratifying in some sense (albeit, if this is right, in a phenomenological sense).

This position will be clarified later in terms of what it means saying that the phenomena are the source of all behaviour. To do this we must enter into an analysis of the nature of meaning and consciousness. Both in natural and artificial settings.

Following a general approach is necessary for the objective of the ASys program of targeting a universal theory of consciousness —in terms of enabling the construction of better autonomous systems— but it is also of maximal relevance when addressing the construction of systems interacting with humans. In order to provide machines suitable for interacting with humans' lives —and most machines are designed to do so— it is necessary to understand this phenomenological grounding for action in humans and also it may be necessary to investigate the possibilities of such a phenomenological stance concerning the realization of machines.

3 ABSTRACT ARCHITECTURE OF A CONSCIOUS MACHINE

Our strategy in the search for a general architecture for consciousness is based in the identification of a set of architectural principles that will guide the definition of reusable design patterns [7]. An early version of these principles was presented in [25]. These principles offer precise but general definitions of some critical concepts in mind theory (like *representation*, *perception*, *action*, *value*, *consciousness*, etc.).

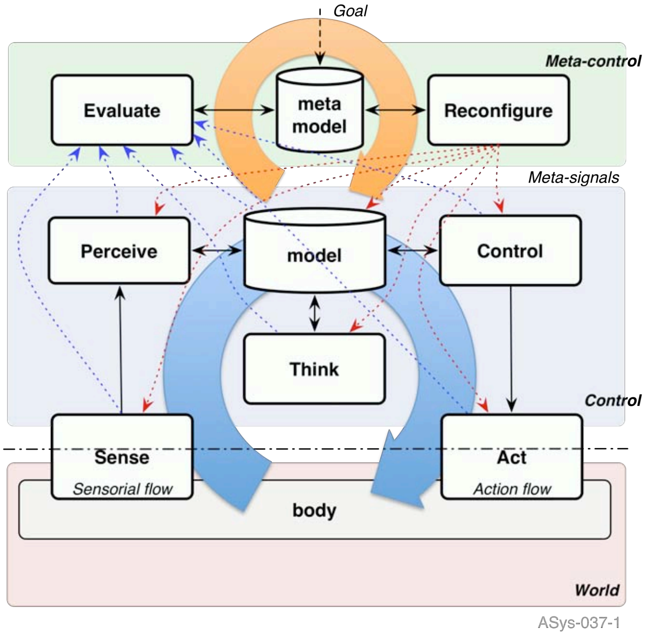

Figure 1. The basic building blocks for a design and realisation of a conscious machine are polymorphic patterns. The figure shows two of the basic patterns used in the definition of the cognitive architecture of reference for general consciousness: EpistemicControlLoop and MetaControl.

The current set of design principles is the following:

1. *A cognitive system builds and exploits models of other systems in their interaction with them*. These models are —obviously— representations. They sustain the realisation of a model-based control architecture. Models are made at multiple levels of resolution and may be aggregated to constitute integrated representations.
2. *An embodied, situated, cognitive system is as good performer as its models are*. The ideal condition is achieving isomorphism in a certain modelling space. It is important to note that models are always abstractions hence defining a modelling space that is inherently different from that of the modelled system.
3. *Except in degenerate cases, maximal timely performance is achieved using predictive models*. What counts for an agent is the value got not only now, but from now on up to a fuzzy time horizon. The depth of the horizon will be dependent of the specific aspect that is anticipated.
4. *Perception is the continuous update of the integrated models used by the agent by means of real-time sensorial information*. Perceiving is hence much more than sensing. Sensing is the mapping of physical estates of the sensed entity into informational states inside the perceiving agent. In a second stage perceptual mechanics updates/creates models to exploit this information. Note that models are necessarily based on a sustaining ontology. This implies that perception suffers model-related ontological blindness.
5. *Agents perceive and act based on multiple integrated, scalable, unified model of task, environment and self*. Model-based control is the core mechanism for action generation. This enables a search for global performance maximisation (obviously bounded by what is known/modelled). Model and action integration may happen at multiple scales.
6. *An aware system is continuously perceiving and computing meaning from the continuously updated models*. Meaning is defined as the partitioning of state-space trajectories in terms of value for the agent. What is different in this proposal for a concept of meaning is that we are considering not only the current state of affairs but the potential future values for the agent.
7. *Models are executed by engines and may be collapsed with them into simpler subsystems*. Model execution leverages models in the obtainment of many classes of data of relevance to the agent: actions, states, causes, means, etc. Model execution is hence necessarily continuous, multiple —forward, backward, means-ends, etc.— and concurrent. In some cases models and engines may be collapsed into a simple, more efficient element. Model-engine collapses are efficiency-exploitability tradeoffs. Collapsed models sacrifice multiple use to gain effectiveness.
8. *Attentional mechanisms allocate both physical and cognitive resources for system perceptive and modelling processes so as to maximize performance*. The bandwidth of the sensory system is enormous and the perceptual task is not easy. The amount of sensed information that may be integrated in the mental models of the agent is bounded by the availability of resources. The allocation of resources to subsets of sensed information is done using cognitive control and also immediate anticipatory valuation (significance feedback).

Note that this implies a primary form of perception before the conscious level.

9. *The agent reconfigures its functional organisation for context-pertinent behaviour using value-driven anticipatory metasignals*. This is the role played by (some) emotional mechanisms [24].
10. *A self-aware system is continuously generating meanings from continuously updated self-models*. The agent perceives and controls itself as it perceives and controls the world. “Self” is the closure of the executing self-model.

These principles are being reified in the form of design patterns (see Figure 1) and implemented using state of the art object-oriented software technologies.

This pattern-based approach enables the formerly stated vision of having both a general approach and the concrete implementations necessary for the diversity of tasks that an agent must address.

In this line of work, Hernández has proposed The Operative Mind (OM) [17] as an architectural framework for development of bespoke systems. This class of architectural reference model —in the line of RCS [1] or CogAff [29]— can be used for engineering systems which implement, as we claim, analogue functional capabilities to those reported —top-down causality, flexible control, integration, informational access, and intrinsic motivation— of biological consciousness. This enables, as a result, improved autonomy and robustness.

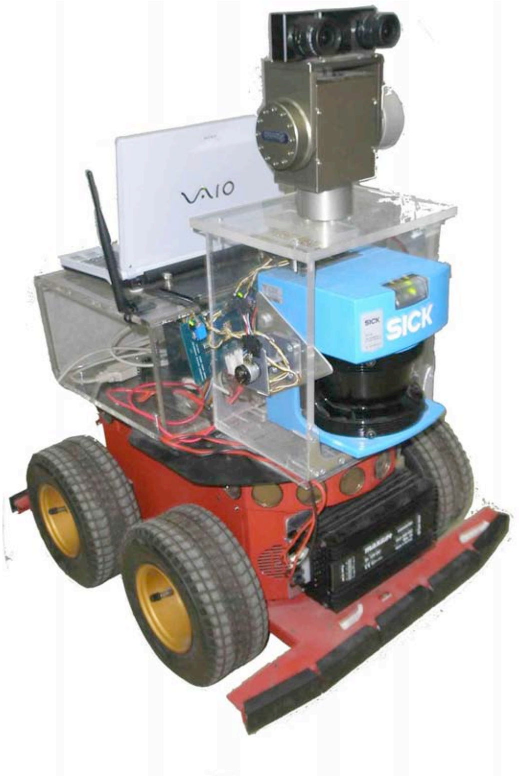

Figure 2. The Higgs robot is the experimental platform used for the deployment of the OM Cognitive Architecture.

Consciousness is implemented on it as a set of services, in an operating system fashion, based on deep modelling of its own control architecture [18], that supervises the adequacy of its structure to the current objectives in the given environment [20] triggering and managing adaptivity mechanisms. This system is being implemented in the control system of an autonomous mobile robot (see Figure 2).

4 MODEL-BASED PREDICTIVE CONTROL AND PHENOMENOLOGY

The architectural model proposed in the above principles is consonant with the model-based control strategies used in technical environments —industrial plants, aircraft, etc. [8].

In model-based predictive control (MBPC), the controller produces the next instantaneous action by i) first projecting a desired trajectory of targets optimised for that goal, ii) then predicting the future consequences of the actions needed to follow that trajectory to obtain precisely an optimised plan of actions, and finally iii) executing only the first action in the plan; then the cycle starts over again.

Notice that for step i) a cost function is used, which is both a model of the task and an evaluation procedure, and for ii) a model of the plant –i.e. system (body) and environment– is employed.

So far, control systems based on advanced techniques such as MBPC contain informational structures and processes that our model could ascribe to access consciousness: they exploit updated models of the plant and evaluate in the view of the predicted future. But as far as the model do not include the system itself –i.e. the controller–, the system is not self-conscious. This implies that there are no phenomenological states concerning the own agent involved.

Now let us suppose that the system/controller includes a model of itself, so it evaluates not only the future environment states given its possible actions, but also its very own possible future states. Then we will have a system that, from sensory information flow, would generate informational structures containing an evaluation of its processing, not only current, but as predicted in the future according to its past.

It is important to note that the evaluation is realised in terms of the value obtained by the agent. In the case of artificial control systems these values are imposed by externally grounded utility functions. In the case of biological systems these utility functions are internal and expressed in terms of what is good and bad for the agent: i.e. its experience. The metaperception of the agent as perceiver sustains the valuation of goodness of states. This may constitute the very substrate of phenomenology: the system, by virtue of the described process, would be *experiencing* that sensory input.

The grounding of experience on model-based metaperception provides an operational understanding of the “what is it like to be” question [22]. To know what is it like to be a bat would require not only the echolocation sensory system but the full perceptual pipeline and the metaperceptual pipeline. We cannot experience being a bat if we don’t meet these requirements, but, however, we can have a deep theory of what it is like to be a bat and hence know “what is it like to be it”.

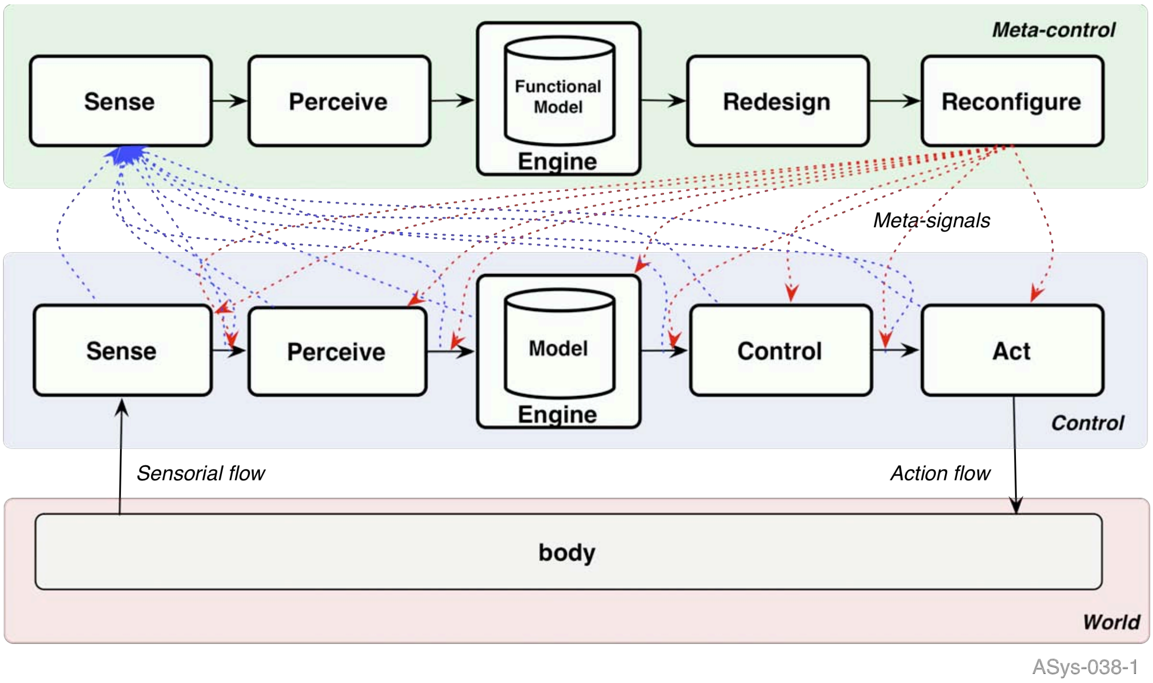

Figure 3. The self-perception, self-configuration meta-loop shares the patterned structure of the EPISTEMICCONTROLLOOP. The meta-level gathers information about the functional organisation of the lower epistemic control loop and may act to change it. The observed/controlled world of the metaloop is a functioning cognitive agent.

Note that the action part of the meta loop shown in Figure 3 shows action modifying the workings of the lower, world-situated loop. The meta-control competences enabled by self perception constitute the active part of emotional mechanisms [24]. In a sense, consciousness, meaning and emotion are stepping-stones in the same road [2].

5 MEANING AND THE FUTURE

In this paper we provide an analysis of 'consciousness" in terms of a precise definition of how an agent "understands" the informational flows entering the agent. This definition of understanding is based in the analysis and evaluation of phenomenological states along potential trajectories in the phase space of the agents.

We propose a rigorous definition of "meaning" in terms of the separation of potential agent trajectories in different value classes —consider that the information flows are a critical resource for trajectory enaction and separation. The values to be computed will not be in the particular space of magnitudes of an external, third person observer but in the magnitudes of relevance to the agent: i.e. the phenomenological ones. This computation requires from the agent an intrinsic capacity for anticipation —including anticipation of phenomenological states.

Note that in this context *phenomenological* is not restricted to the limited interpretation in terms of qualia, but in the broader sense of *phenomenal structure* [30]:

"the phenomenal structure of experience is richly intentional and involves not only sensory ideas and qualities but complex representations [our models] of time, space, cause, body, self, world and the organized structure of the lived reality"

For the reasons stated before, this model —of meaning and consciousness— shall be of applicability both to humans and robots, hence implying a rigorous analysis and definition of phenomenological states —because rigour is necessary if this is going to be built into the robots and not just predicated from some externally observed behaviour.

Clarifying these issues is not only of relevance for robot construction but also for advancing into a general theory of consciousness both operational in the technological side and explanatory in the biological one —e.g. being useful to create safer machines [25] and being able to explain the nature of pain asymbolia [14].

Consider the situation of a system at certain time (now, t_0) where the system must decide what to do based on a certain information it has received (see Figure 4). The system has followed a certain trajectory $x(t)$ in its state space but the future is open concerning the different possibilities for acting (A_a, A_b, A_c). The concrete future trajectory will depend on the concrete action, but will also depend on the concrete state of the world and the agent at t_0. The meaning of a piece of information —about the world or about the agent itself– is the way it partitions the set of possible future trajectories in terms of anticipated phenomenological states.

How is this meaning enacted? By integration of the information received into the model that the agent uses to predict the future and by executing this model in forward time. In a sense, grasping the meaning of some information is leveraging this information in enhancing the prediction of how reality is going to behave.

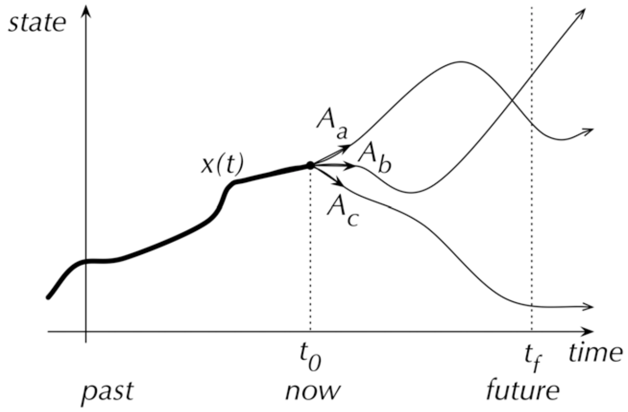

Figure 4. Understanding sensory flows and the derived emotional processes are strongly related to the anticipatory capabilities of the agents.

This interpretation of meaning and consciousness is indeed not new. As Woodbridge said [31] in relation to potential definitions of consciousness [6]: *Professor Bode states the general problem tersely, it seems to me, when he asks, "When an object becomes known, what is present that was not present the moment before ?" I have attempted to answer that question in one word — "meaning."*

Phenomenology goes beyond the experiential qualities of sensed information. Haikonen argues that *qualia are the primary way in which sensory information manifests itself in mind* [16] but in our model this qualitative manifestation is not necessarily primary but may be produced in downstream stages of the perceptual pipeline. What is important for us is not just the qualities of the sensed but the experience of their meaning. As Sloman and Chrisley [29] say, "an experience is constituted partly by the collection of implicitly understood possibilities for change inherent in that experience."

It must be noted that the model proposed is concurrent. This implies that the perceptual pipeline is operating in several percepts at the same time. But due to the integrated nature of the models —principle 5— these pipelines may eventually converge (in non pathological cases). This may imply a reduction of the focus of inner attention to a single percept. This is in line with Dennett's multiple drafts theory of consciousness [11].

6 CONCLUSSIONS: IS HETEROPHENOMENOLOGY A NEED ?

Going back to the analysis done at the very beginning of the paper on the construction of autonomous systems, and after describing the architectural picture of the ASys model of autonomy and consciousness, we reach the conclusion that heterophenomenology is a need.

However, heterophenomenology (phenomenology of others different from oneself) must be understood in a sense a bit different from the initial proposal of the term by Dennett [12] of using verbal reports (and other types of acts) as objective, third-person observations that provide the observer with partial information about the agent's beliefs regarding its own conscious experience.

In this context, building machines that experience, the problem of engineering the right phenomenological mechanism is crucial because it will be the origin of the intrinsic motivations of the agents. We must adopt an heterophenomenological engineering approach in the sense of being able to engineer phenomenologies into machines to match our very own needs [33]. These will not be human phenomenologies but the phenomenologies that when deployed will make the agents pursue our satisfaction.

But for this, we need not only a better understanding of the artificial [27] but of our own consciousness.

7 ACKNOWLEDGEMENTS

We acknowledge the support of the European Commission through Grant *HUMANOBS: Humanoids that learn Socio-communicative Skills by Imitation*.

REFERENCES

[1] James S. Albus, 'Outline of a theory of intelligence', *IEEE Transactions on Systems, Man and Cybernetics*, 21(3), 473–509, (1991).

[2] Yuri I. Alexandrov and Mikko E. Sams, 'Emotion and consciousness: ends of a continuum', *Cognitive Brain Research*, 25, 387 – 405, (2005).

[3] Bernard J. Baars, 'In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness.', *Journal of Consciousness Studies*, 4, 292–309, (1997).

[4] Xabier Barandiarán and Kepa Ruiz-Mirazo, 'Modelling autonomy: Simulating the essence of life and cognition', *Biosystems*, 91(2), 295–304, (2008).

[5] Ned Block, 'On a confusion about the function of consciousness', *Behavioral and Brain Sciences*, 18, 227–247, (1995).

[6] B. H. Bode, 'Some recent definitions of consciousness', *Psychological Review*, 15, 255–264, (1908).

[7] Frank Buschmann, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal, *Pattern Oriented Software Architecture. A System of Patterns*, John Wiley & Sons, Chichester, UK, (1996).

[8] Eduardo F. Camacho and Carlos Bordons, *Model Predictive Control*, Springer, second edn., (2007).

[9] Peter Carruthers, *Phenomenal Consciousness*, Cambridge University Press, Cambridge, (2000).

[10] Antonio Chella, Marcello Frixione, and Salvatore Gaglio, 'A cognitive architecture for robot self-consciousness', *Artificial Intelligence* in Medicine, 44, 147–154, (2008).

[11] Daniel Dennett, *Consciousness Explained*, Penguin, (1991).

[12] D.C. Dennett, 'Whos on first? heterophenomenology explained', *Journal of Consciousness Studies*, 10, 19–30, (2003).

[13] Stanley P. Franklin, 'Building Life-Like 'Conscious' Software Agents', *Artificial Intelligence Communications*, 13, 183–193, (2000).

[14] Nicola Grahek, *Feeling Pain and Being in Pain*, MIT Press, second edn., 2007).

[15] Pentti O. Haikonen, *The Cognitive Approach to Conscious Machines*, Imprint Academic, Exeter, (2003).

[16] Pentti O. Haikonen, ‘Qualia and conscious machines’, *International Journal of Machine Consciousness*, 1(2), 225–234, (2009).

[17] Carlos Hernández, Ignacio López, and Ricardo Sanz, ‘The operative mind: a functional, computational and modelling approach to machine consciousness’, *International Journal of Machine Consciousness*, 1(1), 83–98, (June 2009).

[18] Owen Holland and Ron Goodman, ‘Robots with internal models - a route to machine consciousness?’, *Journal of Consciousness Studies*, 10(4-5), 77–109, (2003).

[19] Lyle N. Long and Troy D. Kelley, ‘The requirements and possibilities of creating conscious systems’, in Proceedings of the *AIAA InfoTech@Aerospace Conference*, Seattle, USA, (April 2009).

[20] Ignacio López, *A Framework for Perception in Autonomous Systems*, Ph.D. dissertation, Departamento de Automática, Universidad Politécnica de Madrid, (May 2007).

[21] Raúl Arrabales and Araceli Sanchis, ‘Applying machine consciousness models in autonomous situated agents’, *Pattern Recognition Letters*, 29(8), 1033–1038, (2008).

[22] Thomas Nagel, ‘What is it like to be a bat?’, *The Philosophical Review*, (October 1974).

[23] Michael Pauen, ‘Emotion, decision, and mental models’, in *Mental Models and the Mind*, eds., Carsten Held, Markus Knauflf, and Gottfried Vosgerau, Elsevier, (2006).

[24] Ricardo Sanz, Carlos Hernández, Jaime Gómez, and Adolfo Hernando, ‘A functional approach to emotion in autonomous systems’, in *Brain Inspired Cognitive Systems 2008*, eds., Amir Hussain, Igor Aleksander, Leslie S. Smith, Allan Kardec Barros, Ron Chrisley, and Vassilis Cutsuridis, volume 657 of Advances in Experimental Medicine and Biology, 249–265, Springer, New York, (2010).

[25] Ricardo Sanz, Ignacio López, and Julita Bermejo-Alonso, ‘A rationale and vision for machine consciousness in complex controllers’, in *Artificial Consciousness*, eds., Antonio Chella and Riccardo Manzotti, Imprint Academic, (2007).

[26] Murray Shanahan, ‘A cognitive architecture that combines internal simulation with a global workspace’, *Consciousness and Cognition*, 15(2), 433–449, (June 2006).

[27] Herbert A. Simon, *The Sciences of the Artificial*, MIT Press, Cambridge,USA, third edn., (1996).

[28] Aaron Sloman, ‘Phenomenal and access consciousness and the ”hard” problem: A view from the designer stance’, *International Journal of Machine Consciousness*, 2(1), 117–169, (2010).

[29] Aaron Sloman and Ron Chrisley, ‘Virtual machines and consciousness’, *Journal of Consciousness Studies*, 10(4-5), 133–172, (April/May 2003).

[30] Robert van Gulick, ‘Consciousness’, in *Stanford Encyclopedia of Philosophy*, Stanford University, (2004).

[31] Frederick J. E. Woodbridge, ‘Conciousness and meaning’, *Psychological Review*, 15(6), 397 – 398, (1908).

[32] Paul Ziff. The feelings of robots. *Analysis* 19(3), January 1959: 64-68. (1959).

[33] Ron Chrisley. Synthetic phenomenology. *International Journal of Machine Consciousness*, 1:53–65. (2009).