International Journal of Machine Consciousness
Vol. 4, No. 2 (2012) 383–399
© World Scientific Publishing Company
DOI: 10.1142/S1793843012400227

CONSCIOUSNESS, ACTION SELECTION, MEANING AND PHENOMENIC ANTICIPATION

RICARDO SANZ*, CARLOS HERNÁNDEZ† and M. G. SÁNCHEZ-ESCRIBANO‡

Autonomous Systems Laboratory,
Universidad Politécnica de Madrid,
José Gutiérrez Abascal 2, 28006 Madrid, Spain
**ricardo.sanz@upm.es*
†*carlos.hernandez@upm.es*
‡*mguadalupe.sanchez@upm.es*

Phenomenal states are generally considered the ultimate sources of intrinsic motivation for autonomous biological agents. In this article, we will address the issue of the necessity of exploiting these states for the design and implementation of robust goal-directed artificial 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 and its very own action possibilities. This abstract model of consciousness and understanding will be based in the analysis and evaluation of phenomenal states along potential future trajectories in the state space of the agents. This implies that a potential strategy to follow in order to build autonomous but still customer-useful systems is to embed them with the particular, *ad hoc* phenomenality that captures the system-external requirements that define the system usefulness from a customer-based, requirements-strict engineering viewpoint.

Keywords: Consciousness; meaning; phenomenology; artificial systems; requirements-driven engineering.

1. Introduction

Machine consciousness research is generally justified as a source of experimentation possibilities with models of human consciousness in order to evaluate their explanatory value [Long and Kelley, 2009; Sun, 2008], but if consciousness has functional value for animals, it is also justifiable in terms of the potential increase of functionality that conscious machines would offer [Sanz, 2007].

Even when there are old arguments against the possibility of machine consciousness,[1] several attempts at realizations of machine consciousness have been made recently [Long and Kelley, 2009]. In some cases, these systems propose a concrete theory of consciousness explicitly addressing artificial agents [Haikonen,

[1] 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."

2003; Chella *et al.*, 2008] but in other cases, the implementations follow psychological or neural theories of human consciousness that were developed by their authors 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 (GWT) of consciousness [Baars, 1997; Arrabales and Sanchís, 2008; Franklin, 2000; Shanahan, 2006].

The machine-based, model testing activities are very valuable efforts that help clarify the many issues surrounding natural consciousness and foster a movement toward 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 toward an ostensive definition of *consciousness* that includes humans and maybe also some animals [Barandiarán and Ruiz-Mirazo, 2008].

However as indicated by Sloman [2010], these multiple efforts may miss the target of a unified theory of consciousness:

> "*...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, fundamental theory of consciousness against which experiments are done and evidence is raised. This is indeed the path followed by the previously mentioned works of Chella, Haikonen, Franklin, Arrabales or Shanahan, taking GWT as this background theory. However, the multifarious character of consciousness is an obvious problem [Block, 1995], which most of the approaches circumvent by focusing on just one aspect of it. Access consciousness seems to be the main target, leaving phenomenality to further clarifications of the hard problem [Chalmers, 1996].

Indeed, Sloman [2010] suggests that the main difficulty that we confront in research on consciousness and machine consciousness is related to this very *polymorphic* nature of the *consciousness* concept. Sloman analysis may seem to imply that trying to tackle several aspects of consciousness — access consciousness, phenomenal consciousness, self-awareness, etc. — in one single shot, in a single model and in a single robot, is hopeless. This program of addressing consciousness as a whole is also hampered by the semantic flaws that some of the conceptions of consciousness suffer when abstracted from specific contexts. However, the general consideration is that while all these consciousness traits may be different *aspects*, they are aspects of a single core phenomenon.

However, Sloman also recognizes 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 Autonomous Systems (ASys) research program. ASys is a program based on general and artificial autonomy. Machine consciousness is just one step: 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.[2]

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

Machines are always built with a purpose [Simon, 1996]. Airplanes are built to carry people across long distances, borers are built to make holes and fans are built to refrigerate. To fulfil their missions in the real world, the machines must be robust. This robustness must be not only physical but also behavioral. In complex machines that perform complex tasks, robust behavior is achieved by means of control systems. These control systems capture the teleological purpose of the machine and drive the body to attain it.

The search for machine autonomy is motivated by several reasons but, in general, responds to the need to make machines that are able to deal with uncertainty and change without the need of human help. Autonomous machines are able to pursue goals in the presence of disturbances [Åström and Murray, 2008]. Machines become autonomous agents pursuing goals.

However, 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*. What goals does an artificial agent pursue? Does it run after its own goals or those of its owner (see Fig. 1). This is very much connected with the value systems of humans and how these drive their behavior [Pauen, 2006]. While goals can be shared and/or transferred, agents can only pursue the goals that are instantiated in their very architecture. The mapping from *external goals*, i.e., those of the owner, to *internal goals*, i.e., those of the agent, may be simple but may be also a very tricky issue. Consider all those machines that go berserk because they continue doing what was specified, regardless of the circumstances (the pump that continues pumping gasoline trough a broken pipe, the oven that continues heating the already cooked chicken, etc.). Machines are sometimes too stubborn in the pursuing of their *mapped* goals.

In this article, we examine phenomenality as the right context for the analysis of motivated action. Phenomenological states are generally considered strong influences or even sources of intrinsic motivation for autonomous biological agents. At

[2] *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).

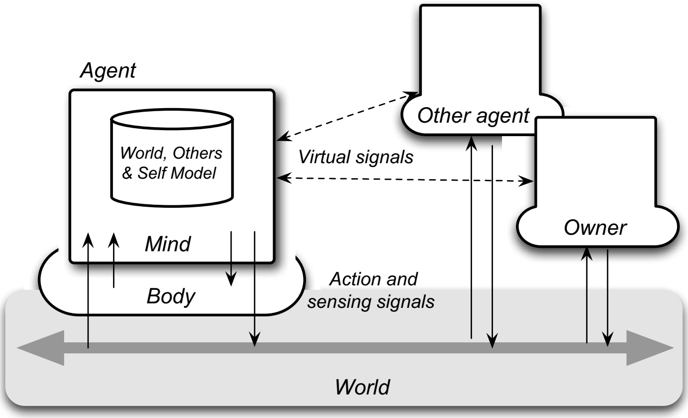

Fig. 1. Agents interact with the world and with other agents. In the case of artificial systems, they act in order to fulfil goals coming from an external owner. The models used by the agent drive its behavior.

the end of the day, what counts for an agent is the array of phenomenal experiences it has had. Agents do what they do in order *to experience the feelings* that they experience doing it. What is relevant for the agent is how the internal changes concerning its perception of the world and of itself impacts its experiential state [Carruthers, 2000]. In a sense, agent actions are just ways of controlling their perceptions [Powers, 2005].

This valuation of phenomenality is not restricted to the agent experiential state in the present but may incorporate extended periods of time in agents able to anticipate experience. To be more precise, for us humans and other anticipatory animals, what counts is the integral, i.e., an accumulated value, of the phenomenal states along the (potentially) lived trajectories — past, present and future. This *valuing of accumulated experiences* 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 behavior shall be gratifying in some way (albeit if this is right, in a phenomenological sense).

This position will be clarified in Sec. 5 in terms of what it means saying that the phenomena are the source of all behavior. To do this, we must enter into an analysis of the nature of meaning and consciousness. This can be done with the desired generality in both 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 that interact 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 may be necessary to investigate the possibilities of such a phenomenological stance concerning the realization of machines.

3. Abstract Architecture of a Conscious Machine

The ASys research program intends the development of universal technology for autonomy. This means that the technology is not being built for a concrete application niche — e.g., mobile robotics — but considers any potential kind of application domain: from robots to mobile phones and from pacemakers to continent-wide electrical networks.

A fundamental component of this technology is the capability of awareness that shall be embedded into the machines to deal with changing worlds and with their own internal mechanics. We intend machine consciousness to provide resilience against external and internal perturbations. We use the tem *self-x* to refer to the capability of a system to *actively act upon itself* in pursuit of some goals [Sanz *et al.*, 2005]. This, in general, requires both self-observation and self-action to implement inner control loops of structural/functional nature [Sanz and López, 2000]. Some technical implementations of concrete *self-x* mechanisms for improving resilience are already widely available, such as adaptive control, fault-tolerant control, autonomic computing, etc. But all of them fall short when considering: (i) the general problem of *self-x* and (ii) agent phenomenality.

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 [Buschmann *et al.*, 1996]. Design patterns are design assets that can be used constructively to generate a complete architecture for an agent. By composing patterns, we can generate architectures that offer the composed functionality offered by each of the patterns.

The pattern-based strategy is rooted in a set of general design principles. An early version of these principles was presented in Sanz *et al.* [2007]. These principles offer precise but general definitions of some critical concepts in mind theory (like *representation*, *perception*, *action*, *value*, *consciousness*, etc.) that are operational in the definition of agent control architectures.

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, *dynamical* representations of other systems. They sustain the realization of a model-based control architecture. Models are made at multiple levels of resolution and may be aggregated to constitute integrated representations. This principle is indeed the fundamental principle behind the ASys vision: cognition is model-based control behavior.

(2) **An embodied, situated, cognitive system is as good a performer as its models are:** The ideal condition for a model-based controller is the achievement of isomorphism in a certain modeling space. It is important to note that models are always abstractions, hence they always define a modeling space that is inherently different from that of the modeled system. There being an isomorphism does not necessarily imply it has to be a complex one, i.e., always

using a complex model. Simple models are usually most valuable — especially when operating in real time — and simplification can go down even to the level of having an immediate, direct coupling in the mode of behavior-based robotics.

(3) **Maximal timely performance is achieved using predictive models:** Using predictive models agents can perform mental time-travel and evaluate future and past states even from a phenomenological perspective. What counts for an agent is the value got not only now, but from now on up to a maybe fuzzy time horizon. The depth of the horizon will be dependent of the specific aspect that is anticipated.

(4) **Models may be multiple, distributed and heterogeneous but shall scale up and be integrable to obtain an effective, system-wide control:** While task-specific isolated models may be useful, the long-term, multi-objective control needs of embodied agents force a drive for model integration and model coherence. Model-based control loops drive not only external action but also the inner control loops and the very model federation mechanics.

(5) **Perception is the continuous update of the models used by the agent by means of real-time sensorial information:** Perceiving is hence much more than sensing [López, 2007]. Sensing is the mapping of the physical states of the sensed entity into informational states inside the perceiving agent. Perception involves a second stage that updates/creates models to exploit this information. Note that models are necessarily based on a sustaining ontology. This implies that perception may suffer model-related ontological blindness. We can see only what we are prepared to see.

(6) **Agents perceive and act based on unified model of task, environment and self:** Model-based control is the core mechanism for action generation. This enables a search for global performance maximization (obviously bounded by what is known/modeled). Model and action integration may happen at multiple scales but they always address the agent and its environment as targets and mediators of its task [Sanz *et al.* 2000].

(7) **An aware system is continuously perceiving and computing meaning from the continuously updated models:** Perception — as model integration — is not sufficient for awareness. Understanding what was perceived is also necessary. Agents shall compute the meaning of what they perceive to be aware and this computation is based on the state of the agent. Meaning is defined as the partitioning of state-space trajectories in terms of the value they have for the agent. What is different in this proposal for a concept of meaning is that it considers not only the current state of affairs but the potential future values for the agent.

(8) **Models are exploited by engines and may be collapsed with them into simpler subsystems:** Model exploitation — usually 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 simple, more efficient components. Model-engine collapses are efficiency-exploitability trade-offs. Collapsed models sacrifice the explicitness that enables multiple uses to gain effectiveness — in time and in space.

(9) **Attentional mechanisms allocate both physical and cognitive resources for system perceptive and modeling processes so as to maximize task-oriented 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 the processing resources. The allocation of resources to subsets of sensed information is done using cognitive conscious and unconscious control and also immediate anticipatory valuation (significance feedback) [Herrera *et al.*, 2012]. Note that this implies a primary form of perception before the conscious level [Hernández *et al.*, 2008].

(10) **The agent reconfigures its functional organization for context-pertinent behavior using value-driven anticipatory metasignals:** This is the role played by (some) emotional mechanisms [Sanz *et al.*, 2010; Sanz, 2011; Herrera and Sanz, 2013; Sánchez-Escribano and Sanz, 2012].

(11) **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. Interoceptive and proprioceptive signals are

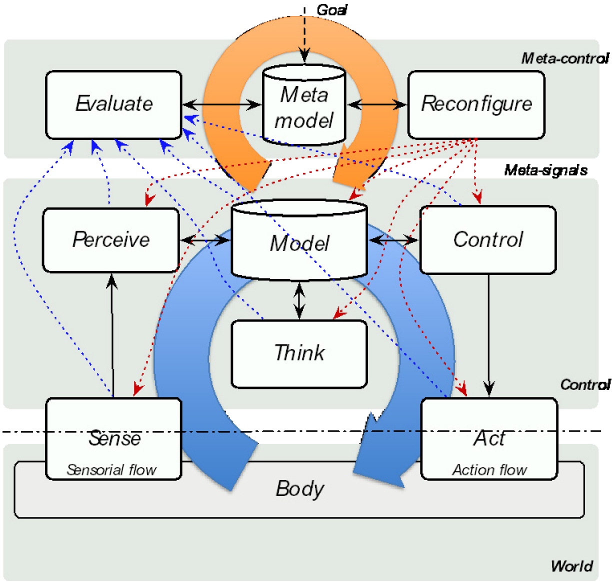

Fig. 2. The basic building blocks for a design and realization of a conscious machine are polymorphic patterns. The figure shows two of the basic patterns used in the definition of the cognitive architecture of reference for general consciousness: Epistemic-Control-Loop and Meta-Control.

injected into self-models of the agent. "The self" is the control-closure of the executing self-model [Hernández and Sanz, 2012].

These principles are being reified in the form of design patterns (see Fig. 2) and implemented using state-of-the-art object-oriented software technologies. An example pattern is the Epistemic Control Loop. This loop captures the first ASys principle in the form of a model-based control loop. Knowledge is equated to models and used to drive agent actions.

4. From Abstractions to Real Systems

The ASys pattern-based approach to the engineering of autonomous systems enables the formerly stated vision of having: (i) a general approach and (ii) the concrete implementations necessary for the diversity of tasks that an agent must address.

In this line of work, Hernández *et al.* [2009] have proposed the Operative Mind (OM) as an architectural framework for development of bespoke systems. This class of architectural reference model — in the line of RCS [Albus, 1991] or CogAff [Sloman and Chrisley, 2003] — can be used for engineering systems which implement, as we claim, functional capabilities that are analogous to those reported — top-down causality, flexible control, integration, informational access, and

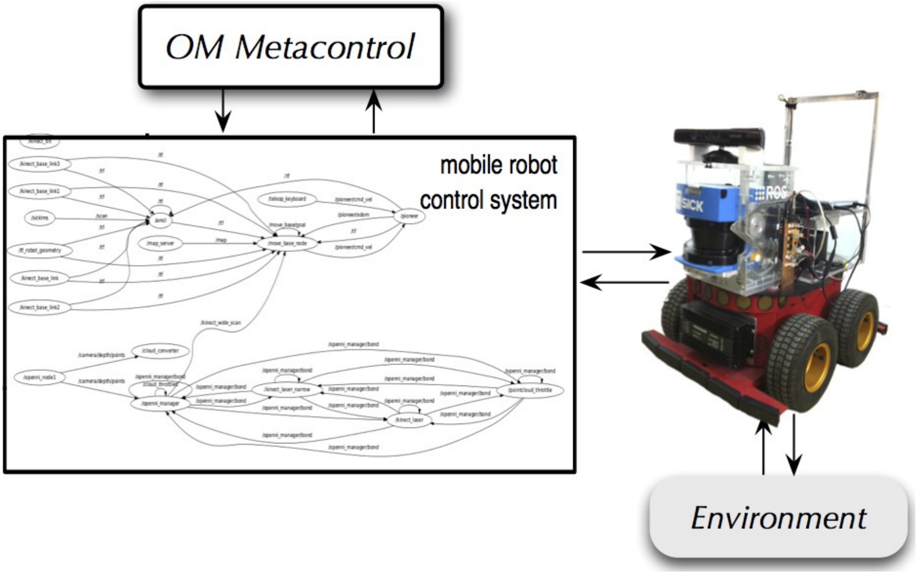

Fig. 3. The Higgs robot is the experimental platform used for the deployment of the OM Cognitive Architecture. The OM meta-control subsystem perceives the functional state-of-the-mobile robot's control system, and adapts it through reconfiguration if its structure is not rendering the required behavior. The OM metacontroller reifies the OM architecture making use of the Epistemic-Control-Loop and Meta-Control patterns.

intrinsic motivation — of biological consciousness. This enables, as a result, improved autonomy and robustness.

OM proposes that machine consciousness could be implemented as a set of infrastructural services, in an operating system fashion, based on deep modeling of the control system's own architecture [Holland and Goodman, 2003], which are provided by a meta-control subsystem that supervises the adequacy of the system's structure to the current objectives in the given environment [López, 2007] triggering and managing adaptivity mechanisms. This system is being implemented in the control system of an autonomous mobile robot (see Fig. 3).

5. 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. [Camacho and Bordons, 2007]. For example, in model-based predictive control (MBPC), the controller produces the next instantaneous action by:

(i) First, projecting a desired trajectory of targets optimized for the system goals;
(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 (a way of addressing the goal mapping problem mentioned before), 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 framework could easily ascribe to access consciousness: they exploit updated models of the plant and evaluate them in the view of the predicted future. But insofar as the model does not include the system itself — i.e., the controller — the system is not self-conscious.

This analysis implies also that if there is no self-model, then there are no phenomenal states concerning the agent itself involved in the perception and decision-making process (cf. Metzinger's [2003] phenomenal self-model). Note that the reverse implication does not hold: a non-self-conscious agent can still have a model of itself when the lack of consciousness is due to the failure of the model exploitation engine.

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 to be realized in terms of the value obtained by the agent — past, present and, more importantly, future. 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.[3] The meta-perception 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 meta-perception provides an operational understanding of the "what is it like to be" question [Nagel, 1974]. To know what is it like to be a bat would require not only the echo location sensory system but the full perceptual pipeline and the meta-perceptual pipeline — including the world and self-models and meaning generation systems. We cannot *experience* being a bat if we do not 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".

Note that the action part of the meta-loop shown in Fig. 4 shows actions modifying the workings of the lower, world-situated loop. The meta-control competences enabled by self-perception constitute the active part of emotional mechanisms [Sanz *et al.*, 2010]. In a sense, consciousness, meaning and emotion are stepping-stones in the same road [Alexandrov and Sams, 2005].

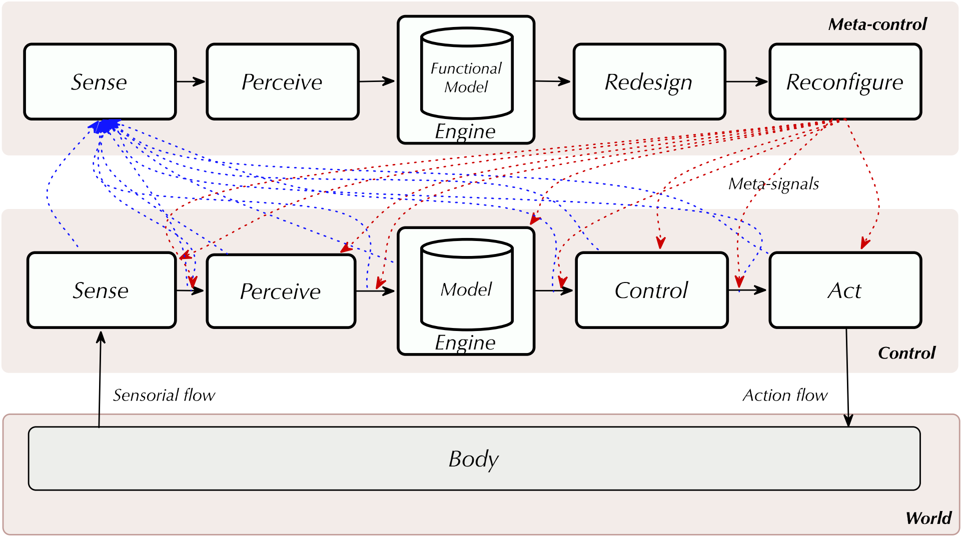

Fig. 4. Meta-control is a self-perception, self-configuration control loop that shares the patterned structure of the Epistemic-Control-Loop. The meta-level gathers information about the functional organization of the lower epistemic control loop and acts to change it. The observed/controlled world of the meta-loop is a functioning cognitive agent.

[3] This is shaped evolutionarily in biological systems hence implying *goodness* and *badness* also for the species as a whole, not only for the individuals.

This meta-layered approach is convergent with several other theoretical models of consciousness, including higher-order theories and supramodular interaction theory. In particular, this last one is specially relevant due its architectural nature, proposing that phenomenal states play an essential role in "*permitting interactions among supramodular response systems — agentic, independent, multimodal, information-processing structures defined by their concerns*" [Morsella, 2005].

6. Meaning and the Future

In this article, we intended to provide an analysis of "consciousness" in terms of a precise definition of how an agent "understands" the informational flows entering that agent [Kuipers, 2005]. This definition of understanding is based in the analysis and evaluation of phenomenal states along alternative trajectories in the state space of the agents.

We propose a rigorous definition of "meaning" in terms of the separation of state-space 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 phenomenal ones. This computation requires from the agent an intrinsic capacity for anticipation — including anticipation of phenomenal states. The predictive capability affects the current decision-making but does not fully shape the future, because the prediction-decision process will continue in the immediate future. In a sense, the future is always open to change and only the present is determined by the confluence of the current state and the anticipation.

We should be aware, however, that in this context "phenomenal" is not restricted to a limited interpretation in terms of qualia, but in the broader sense of phenomenal structure [van Gulick, 2004]:

> "*...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 rigor is necessary if this is going to be built into the robots and not just predicated from some externally observed behavior.

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 [Sanz *et al.*, 2007] or being able to explain the nature of pain asymbolia [Grahek, 2007].

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 Fig. 5). The

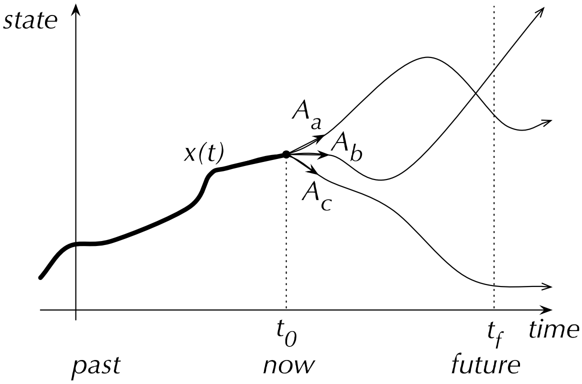

Fig. 5. Understanding sensory flows and the derived emotional processes are strongly related to the anticipatory capabilities of the agents. Action decisions (A_a, A_b, A_c) to be taken at present time (t_0) are influenced by the phenomenal anticipation along potential future trajectories of the system.

system is following a certain trajectory $x(t)$ in its state space but the future is still open concerning the different possibilities for acting (A_a, A_b, A_c).[4]

The concrete future trajectory of the agent will depend on the concrete action selection done at t_0, but will also depend on the concrete state of the world and the agent at t_0. The agent anticipates the future based on the knowledge that its models possess.

The meaning of any 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. When the agent receives the information, it is integrated into the models used to anticipate the future and used to evaluate it. This is in close relation to Gibsonian affordances, where the agent perceives things in terms of interaction and what can happen with them.

In summary, meaning is 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.

This interpretation of meaning and consciousness is indeed not new. As Woodbridge [1908] said in relation to the possibility of precise definitions of consciousness by Bode [1908]:

> "*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.'*"

[4] Note that this does not mean that we endorse a free will theory of agent behavior. This model is fully deterministic, but given the scope of the analysis (just the agent) not all causes are taken into account and hence alternative trajectories may be considered.

Phenomenology goes beyond the experiential qualities of sensed information. Haikonen [2009] argues that *qualia are the primary way in which sensory information manifests itself in mind* 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 the model presented here is not just the qualities of the sensed but the experience of their meaning. As Sloman and Chrisley [2003] 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 4, these pipelines shall eventually converge (in non-pathological cases) to a common multiobjective control strategy [Sánchez-Escribano and Sanz, 2012]. This may imply a reduction of the focus of inner attention to a single percept or a simple percept bundle. This is in line with Dennett's [1991] multiple drafts theory of consciousness.

The perceptual processes are also happening concurrently and concurrently with the action ones, and all of them may happen at different abstraction and meta-levels. This produces a complex tangle of activities that make difficult the analysis of agent behavior in terms of simple mechanisms (see Fig. 6).

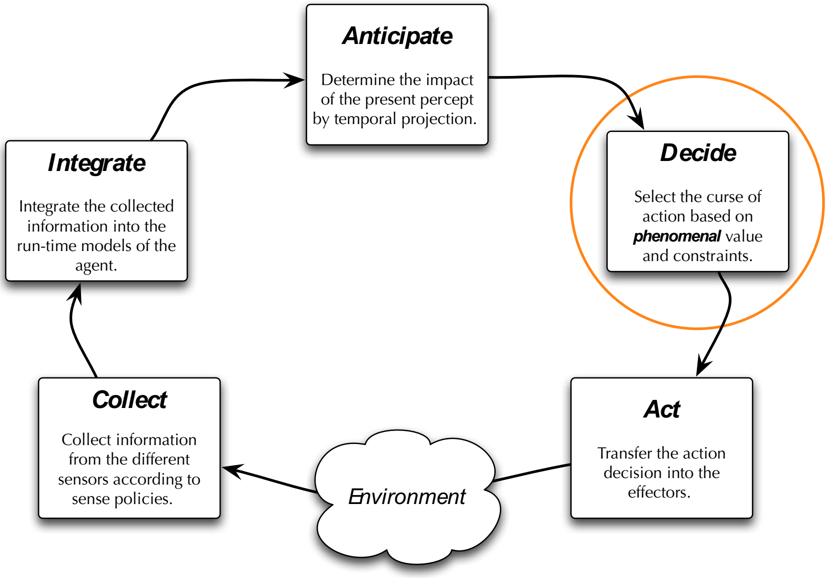
Anticipate
Determine the impact of the present percept by temporal projection.
Decide
Select the curse of action based on ***phenomenal*** value and constraints.
Act
Transfer the action decision into the effectors.
Environment
Collect
Collect information from the different sensors according to sense policies.
Integrate
Integrate the collected information into the run-time models of the agent.

Fig. 6. The activities performed by the epistemic control loop include action selection — the Decide box — that shall be based on the analysis of future phenomenal states. For artificial agents, i.e., agents built with a purpose, the phenomenal value to be computed shall be that of the owner; this has complex implications for artificial agent architecture.

7. Conclusions: Is Heterophenomenology a Need?

Going back to the analysis done at the very beginning of the article on the construction of autonomous systems, and after describing the architectural picture of the ASys model of autonomy and consciousness presented before, we reach the conclusion that *synthetic heterophenomenology* is necessary [Chrisley, 2008].

However, heterophenomenology (phenomenology of others different from oneself) must be understood rather differently from the initial usage of the term by Dennett [2003] 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, i.e., building *autonomous* machines that provide certain services to owners, the problem of engineering the right phenomenic mechanism for them is hence *absolutely crucial* because it will be the origin of the intrinsic motivations of the agents. Autonomy requires a strong focus on machine phenomenality. We must adopt an heterophenomenological engineering approach in the sense of being able to engineer phenomenalities into machines to match our very own needs [Chrisley, 2009]. These will not necessarily be human phenomenalities but the phenomenalities that when deployed will make the agents pursue our satisfaction. Machines shall make their decisions based on predictions of what will be our experiences and, due to the intrinsic motivational mechanics of autonomous agents, these shall be mapped into theirs: artificial — in Simon's sense — autonomous machines shall feel pleasure based on a projection of ours.

There is a problem, however, concerning the practical realization of the phenomenal engines of machines. The appearance of phenomenality in animals may be a non-localized, emergent phenomenon affecting a large number of brain subsystems without a clear neural correlate of consciousness [Koch, 2004; Metzinger, 2000; Morsella *et al.*, 2010]. Robot phenomenality may be created by replicating the functional organization of brains and associated systems. This may require an enormous effort — consider for example those proposed large-scale brain replication projects[5] — but can eventually produce a machine with its own phenomenality.

But for useful artificial systems that pursue owner-centric objectives, there is a need of anticipating not the agent experiences but the owner ones. If phenomenality is not modularized but emergent, this may require the replication of the whole mental architecture of the owner inside the machine. This may be indeed what happens in biological agents and what is being addressed in the theory and the simulation theories of mind [Michlmayr, 2002]. Machines may require deep theories of mind of their owners.

But for all this, we need not only a better understanding of the nature of the artificial [Simon, 1996] but of our very own consciousness. This happening, phenomenology will render a solid scientific foundation to improve autonomous machines' design and better tailor them to human needs.

[5] See, for example, the Human Brain Project: http://www.humanbrainproject.eu/.

Acknowledgment

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

References

Albus, J. S. [1991] "Outline of a theory of intelligence," *IEEE Transact. Syst. Man Cybernetics* **21**(3), 473–509.

Alexandrov, Y. I. and Sams, M. E. [2005] "Emotion and consciousness: Ends of a continuum," *Cogn. Brain Res.* **25**, 387–405.

Arrabales, R. and Sanchís, A. [2008] "Applying machine consciousness models in autonomous situated agents," *Pattern Recogn. Lett.* **29**(8), 1033–1038.

Åström, K. J. and Murray, R. M. [2008] *Feedback Systems: An Introduction for Scientists and Engineers* (Princeton University Press, Princeton).

Baars, B. J. [1997] "In the theatre of consciousness. Global Workspace Theory, a rigorous scientific theory of consciousness," *J. Conscious. Stud.* **4**, 292–309.

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

Block, N. [1995] "On a confusion about the function of consciousness," *Behav. Brain Sci.* **18**, 227–247.

Block, N. [2007] *Consciousness, Function, and Representation.* Collected Papers, Vol. 1 (MIT Press, Cambridge, MA).

Bode, B. H. [1908] "Some recent definitions of consciousness," *Psychol. Rev.* **15**, 255–264.

Buschmann, F., Meunier, R., Rohnert, H. Sommerlad, P. and Stal, M. [1996] *Pattern Oriented Software Architecture. A System of Patterns* (John Wiley & Sons, Chichester, UK).

Camacho, E. F. and Bordons, C. [2007] *Model Predictive Control*, 2nd edition (Springer, Berlin).

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

Chalmers, D. J. [1996] *The Conscious Mind. Philosophy of Mind* (Oxford University Press, Oxford, New York).

Chella, A., Frixione, M. and Gaglio, S. [2008] "A cognitive architecture for robot self-consciousness," *Artif. Intell. Med.* **44**, 147–154.

Chrisley, R. [2008] "Philosophical foundations of artificial consciousness," *Artif. Intell. Med.* **44**(2), 119–137.

Chrisley, R. [2009] "Synthetic phenomenology," *Int. J. Mach. Conscious.* **1**, 53–65.

Dennett, D. C. [1991] *Consciousness Explained* (Penguin, New York).

Dennett, D. C. [2003] "Who's on first? Heterophenomenology explained," *J. Conscious. Stud.* **10**, 19–30.

Franklin, S. P. [2000] "Building life-like 'conscious' software agents," *Artif. Intell. Commun.* **13**, 183–193.

Grahek, N. [2007] *Feeling Pain and Being in Pain*, 2nd edition (MIT Press, Cambridge).

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

Haikonen, P. O. [2009] "Qualia and conscious machines," *Int. J. Mach. Conscious.* **1**(2), 225–234.

Hernández, C., Sanz, R. and López, I. [2008] "Attention and consciousness in cognitive systems," *ESF-JSPS Conf. Series for Young Researchers: Cognitive Robotics*, ESF-JSPS.

Hernández, C., López, I. and R. Sanz [2009] "The operative mind: A functional, computational and modelling approach to machine consciousness," *Int. J. Mach. Conscious.* **1**(1), 83–98.

Hernández, C. and Sanz, R. [2012] "Three patterns for autonomous systems," Technical Note, Autonomous Systems Laboratory ASLab, Universidad Politécnica de Madrid.

Herrera, C., Sanchez, G. and Sanz, R. [2012] "The morphofunctional approach to emotion modeling in robotics", *Adaptive Behavior* **20**(5), 388–404.

Herrera, C. and Sanz, R. [2013] "Emotion as morphofunctionality," *Artif. Life*, **19**(1).

Holland, O. and Goodman, R. [2003] "Robots with internal models — A route to machine consciousness?" *J. Conscious. Stud.* **10**(4–5), 77–109.

Kuipers, B. [2005] "Consciousness: Drinking from the firehose of experience," *Proc. AAAI Conference on Artificial Intelligence*, Vol. 20 (AAAI Press, Menlo Park, USA), pp. 1298–1305.

Long, L. N. and Kelley, T. D. [2009] "The requirements and possibilities of creating conscious systems," *Proc. AIAA InfoTech@Aerospace Conf.*, Seattle, USA.

López, I. [2007] *A Framework for Perception in Autonomous Systems*, Ph.D. Dissertation, Departamento de Automática, Universidad Politécnica de Madrid.

Metzinger, T. [2000] *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (MIT Press, Cambridge, MA).

Metzinger, T. [2003] *Being No One: The Self-Model Theory of Subjectivity* (MIT Press, Cambridge, MA).

Michlmayr, M. [2002] *Simulation Theory Versus Theory Theory: Theories Concerning the Ability to Read Minds.* Master's thesis, Leopold-Franzens-Universität Innsbruck.

Morsella, E. [2005] "The function of phenomenal states: Supramodular interaction theory," *Psychol. Rev.* **112**(4), 1000–1021.

Morsella, E., Krieger, S. C. and Bargh, J. A. [2010] "Minimal neuroanatomy for a conscious brain: Homing in on the networks constituting consciousness," *Neural Networks* **23**(1), 14–15.

Nagel, T. [1974] "What is it like to be a bat?" *Philos. Rev.* **83**(4), 435–450.

Pauen, M. [2006] "Emotion, decision, and mental models," in *Mental Models and the Mind*, eds. Held, C., Knauflf, M. and Vosgerau G. (Elsevier, Amsterdam).

Powers, W. T. [2005] *Behavior: The Control of Perception*, 2nd edition (Benchmark Publications, New Canaan, CT).

Sánchez-Escribano, M. G. and Sanz, R. [2012] "Value by architectural transversality, emotion and consciousness," *16th Meeting of the Association for the Scientific Study of Consciousness*, Brighton, UK.

Sanz, R. and López, I. [2000] "Minds, MIPS and structural feedback," in *Performance Metrics for Intelligent Systems, PerMIS '2000*, Gaithersburg, USA.

Sanz, R., Matía, F. and Galán, S. [2000] "Fridges, elephants and the meaning of autonomy and intelligence," *IEEE Int. Symp. Intelligent Control, ISIC'2000*, Patras, Greece.

Sanz, R., López, I., Bermejo-Alonso, J., Chinchilla, R. and Conde, R. [2005] "Self-x: The control within," *Proc. IFAC World Congress 2005.*

Sanz, R., López, I. and Bermejo-Alonso, J. [2007] "A rationale and vision for machine consciousness in complex controllers," in *Artificial Consciousness*, eds. Chella, A. and Manzotti, R. (Imprint Academic, UK), pp. 141–155.

Sanz, R., Sanchez-Escribano, G and Herrera, C. [2011] "A model of emotion as patterned metacontrol," in *Proceedings of Biologically Inspired Cognitive Architectures*, Washington DC, USA.

Sanz, R., Hernández, C., Gómez, J. and Hernando, A. [2010] "A functional approach to emotion in autonomous systems," in *Brain Inspired Cognitive Systems 2008*, eds. Hussain, A. *et al.* (Springer, New York), pp. 249–265.

Shanahan, M. [2006] "A cognitive architecture that combines internal simulation with a global workspace," *Conscious. Cogn.* **15**(2), 433–449.

Simon, H. A. [1996] *The Sciences of the Artificial*, 3rd edition (MIT Press, Cambridge, USA).

Sloman, A. [2010] "Phenomenal and access consciousness and the "hard" problem: A view from the designer stance," *Int. J. Mach. Conscious.* **2**(1), 117–169.

Sloman, A. and Chrisley, R. [2003] "Virtual machines and consciousness," *J. Conscious. Stud.* **10**(4–5), 133–172.

Sun, R. (ed.) [2008] *The Cambridge Handbook of Computational Psychology* (Cambridge University Press, New York).

van Gulick, R. [2004] "Consciousness," in Zalta, Edward N. (ed.), *Stanford Encyclopedia of Philosophy* (Stanford University, Stanford).

Woodbridge, F. J. E. [1908] "Conciousness and meaning," *Psychol. Rev.* **15**(6), 397–398.

Ziff, P. [1959] "The feelings of robots," *Analysis* **19**(3), 64–68.