THE AUTONOMOUS SYSTEMS LABORATORY
UNIVERSIDAD POLITÉCNICA DE MADRID

Principles for Consciousness in Integrated Cognitive Control

Published in Neural Networks Vol. 20, Issue 9, Nov. 2007

Ricardo Sanz, Ignacio López,
Manuel Rodríguez and Carlos Hernández

ASLab A-2007-011 v 1.0 Final
September 2007

Abstract

In this article we will argue that given certain conditions for the evolution of biological controllers, these will necessarily evolve in the direction of incorporating consciousness capabilities. We will also see what are the necessary mechanics for the provision of these capabilities and extrapolate this vision to the world of artificial systems postulating seven design principles for conscious systems. This article was published in the journal *Neural Networks* special issue on brain and consciousness (Sanz et al., 2007).

1 A Bit of Context

In this our excursion into the scientific worlds of awareness so germane to the world of humanities[1], we observe with surprise that there is still a widely extended perception of *consciousness as epiphenomenon*, which, while mainly rooted in philosophical analyses, is also apparently supported by real, tangible experiments in well controlled conditions (see for example Libet et al. (1982); Varela (1971); Pockett (2004); Dennet (1991)). Hence, we may wonder why engineers should be interested in such a phenomenon that is not yet not only full understood but, somehow, even fully accepted.

[1] Humanities in Snow's sense Snow (1969).

In this paper we will argue for a precise interpretation of consciousness —based on controller mechanics— that renders it not only not epiphenomenal but fully functional. Even more, this interpretation leads to the conclusion that consciousness necessarily emerges from certain, not excessively complex, circumstances in the dwelling of cognitive agents.

A characterisation of cognitive control will be needed as a base support of this argument; and from this initial relatively simple setting, the unavoidable arrow of evolution will render entities that are not only conscious but also necessarily self-conscious.

This analysis will provide a stance for the analysis of the phenomenon of consciousness in cognitive agents that is full in-line with fashionable buzzwords like *situatedness* and *embodiment*.

In the case of technical systems, evolutionary pressure also operates in their evolution. Not at the level of individual machines but at the human-mediated level of product lines and product families (individual machines generally lacking the necessary replicatory capacities for selfish gene evolution). This implies that, sooner or later, if the initial conditions hold in this context, consciousness will be a necessarly appearing trait of sophisticated machines. This is where we are: identifying the core mechanics and application constraints for the realisation of consciousness capabilitites in next generation technical systems. This will imply, necessarily, the sound characterisation of the expected benefits from making a machine conscious.

2 The Modelling Brain

2.1 The modelling principle

One of the central issues proposed by the research community is the question of existence of *general principles for cognitive systems* and of consciousness in particular Aleksander and Dunmall (2003). These are for example the topics of discussion formulated by Taylor in a proposal for a special session on ICANN 2007:

- General principles for cognitive systems;
- The pros and cons of embodiment for cognitive systems;
- The desirability or otherwise of guidance from the brain;
- Specific cognitive system designs and their powers;
- Embodied cognitive systems in robot platforms and demonstrations;
- The best future pathways for development of cognitive systems;

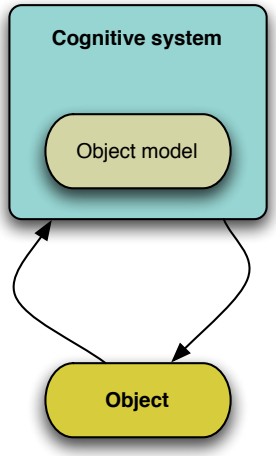

Figure 1: The cognitive relations of a system with an object are mediated by a model of the object.

Proposing some cognitive principles up to the level of consciousness will be the objective of this paper. Let's start with a first one on the nature of cognition:

> **Principle 1:** ***Model-based cognition*** — *A system is said to be cognitive if it exploits models of other systems in their interaction with them.*

This principle in practice equates knowlegde with models, bypassing the problems derived from the conventional epistemological interpretation of knowledge as *justified true belief* Gettier (1963) and embracing a Dretskean interpretation where justification and truth are precisely defined in terms of a strict modeling relation Rosen (1985)[2]. Obviously, this principle takes us to the broadly debated interpretation of cognition as centered around representation, but with a tint; that of the predictive and postdictive capabilities derived from the execution of such a model.

In what follows, just to avoid confusion, we will try to reserve the use of the term *system* for the cognitive system (unless splicitly stated otherwise) and use the term *object* for the system that the cognitive system interacts with (even when in some cases this one may be also cognitive).

Obviously, that the mind uses models is not a new theory. The model-based theory of mind can be traced back in many disciplines and the topic of mental models have been a classic approach to the study of mind Craik (1943); Gentner and Stevens (1983) but this has just had an aura of methaphorical argumentation Johnson (1987) because of the lack of formalisation of the concept of model and

[2]The truth of a piece of information included into a model is not just its fitness into the model —*e.g.* a perspective held by social constructivists— but the in terms of the establishment of isomorphims between the model and the modelled.

the less than rigorous approach to the study of its use in the generation of mental activity.

Closer approaches are for example the emulation theory of representation of Grush Grush (1995) or the model-based sensory-motor integration theory of Wolpert Wolpert et al. (1995). Grush proposed the similar idea that the brain represents external-to-mind things, such as the body and the environment, by constructing, maintaining, and using models of them. Wolpert addresses the hypothesis that the central nervous system internally models and simulates the dynamic behavior of the motor system in planning, control, and learning.

We think that we can go beyond using the concept of *model-based-mind* as metahphor or as *de facto* contingent realizations found in biological brains to the more strong claim that minds are *necessarily* model-based and that evolutionary pressure on them will *necessarily* lead to consciousness. This article is just one step in this direction.

2.2 On models

This definition of cognition as model-based behavior many sound too strict to be of general applicability; in particular it seems not fitting simple cognitive processes (*e.g.* it seems that we can have a stimulus input without having a model of it). However, if we carefully analise these processes we will find isomorphisms between information structures in the system's processes —*e.g.* a sense— and the external reality —the sensed— that are *necessary* for the process to be succesful.

These information structures may be explicit and directly identifiable in their isomorphisms or may be exteemely difficult to tell apart. Models will have many forms and in many cases they may even be fully integrated —collapsed— into the very mechanisms that exploit them. The model information in this case is captured in the very structure of the cognitive process. Reading an *effective* cognitive system tells us a lot about its surounding reality.

The discussion of what is a the proper charaterisation of the concept of model is also very old and plenty of clever insights as that one of George Box: "Essentially, all models are wrong but some are useful" Box and Draper (1987). Is this model usefulness what gives adaptive value to cognition as demosntrated by Conant Conant and Ashby (1970).

There are plenty of references on modelling theory, mostly centered in the domain of simulation Cellier (1991); Zeigler et al. (2000) but it is more relevant for the vision defended here the perspective from the domains of systems theory Klir (2001) and theoretical biology Rosen (1993, 1991).

This last gives us a definition of model in terms of a *modelling relation* that fits the perspective defended in this article: a system A is in a modelling relation with another system B —*i.e.* is a model of it— if the entailments in model A can be

mapped to entailments in model B. In the case of cognitive systems, model A will be abstract and stored in the mind *or the body* of the cognitive agent and system B will be part of its surrouding reality.

We must bear in mind, however, that models may vary widey in terms of purpose, detail, completitude, implementation, etc. A model will represent only those object traits that are relevant for the purpose of the model and this representation may be not only not explicit, but fully fused with the model exploitation mechanism.

2.3 Relations with other traits

Principle 1 grounds some common conceptions about cognitive systems; obviously the most important is the question of *representation*. A cognitive system —by definition of cognition— necessarily represents other systems. Even more, these representations must have deep isomorphisms with the represented objects so the cognitive system can exploit formal entailments in its models to compute entailments in the modelled object in order to maximise the utility of the interaction (more on this in section 3). Paraphrasing what Conant and Ashby clearly stated Conant and Ashby (1970) —every good regulator must contain a model of the system it is controlling— we can say that every well performing cognitive system must contain a model of the objects it is interacting with.

Many other core issues of cognitive systems are addressed by Principle 1. Two quite fashionable these days are the questions of *situatedness* —cognition is necessarily interactive with an external world— and *embodiment* —the necessary separation of the agent body from the rest as defined by the interaction boundary. Both are duly addressed by the modeling perspective of Principle 1 even when they are not as necessarily crisp as they may appear to roboticists because the model can obviosly represent uncertainty and vagueness, hence being able to handle even blurred bodies and fuzzy situations.

2.4 On model generation

Other so-called cognitive traits are left out of this picture of cognitive systems.

Model-based —cognitive— systems need not necessarily be *learning* systems —even while learning will be a very common procedure for model generation. A cognitive system may operate using a static model —coming from any source— as long as it is considered valid. *i.e.* as long as the *modeling relation* with the external object still holds.

Obviously, from the consideration of how the cognitive system becomes cognitive or maintains its cognitive capability learning becomes crucial. Somehow the models must be put there, in the mind of the cognitive system. In general —not

just in the case of biosystems— the core infrastructures for model construction fall in three categories:

Built-ins: In the sense described by Conant and Ashby Conant and Ashby (1970), our feeding, homeostatic and kinestetic mechanisms contain models of the surounding reality (*e.g.* genes codifying chemical receptors for the nose).

Learned: The very subject matter of learning from experience.

Cultural: The well known topic of memetics Dawkins (1976); Blackmore (1999) or —more visually shocking— of Trinity "learning" helicopter piloting expertise in Wachowskys' *Matrix*. [3]

The learning and cultural mechanisms have the extremely interesting property of being *open ended*. In particular, cultural model transmision is a form of extended learning, where the cognitive system downloads models learned by others hence reaching levels of model complexity and perfection that are impossible for an isolated agent[4].

In biological systems, the substrate for learning is mostly neural tissue. Neural networks are universal approximators that can be tuned to model any concrete object or objects+relations set. This property of universal approximation combined with the potential for unsupervised learning make the neural soup a perfect candidate for model boostraping and continuous tuning. The neural net is an universal approximator; the neural tissue organised as brain is an universal modeller.

These are also the properties that are sought in the field of artificial neural networks. It is not necessary to recall here the ample capacities that neural networks —both artificial and natural— have shown concerning model learning. We may wonder to what extent model learning of an external reality can be equated to the advances in modeling external realities demonstrated in the so called hard-sciences (deep, first principles models).

What is philosophically interesting of this process of scientific model construction is the fact that reality seems to have a mathematical-relational stucture that enables the distillation of progressively precise models in closed analytical forms Wigner (1960).

We may think that culturally learnt first principles models[5] are better than neural network approximative modelling[6]; there are cases of exact convergence of both modelling approaches but there are also cases where the mathematical shape of the principles limits their applicability to certain classes of systems.

[3] Supervised learning may be considered an hybrid of cultural and learned processes.
[4] Indeed this is, plainly, the phenomenon of science.
[5] Only geniuses do incorporate first principles models by autonomous learning.
[6] A similar problem to that of having symbolic representations in neural tissue.

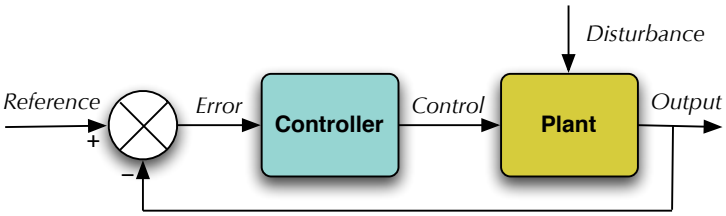

Figure 2: Feedback controllers measure the difference (error) between what we want (reference) and what we have (output) and make corrections (control) based on this difference.

For example, in the field of model creation for control purposes, artificial neural networks have been compared favourably, in certain settings, with first principles models in the implementation of nonlinear multivariable predictive control Henriques et al. (2002). This neural network approach uses a recurrent Elman network for capturing the plant's dynamics, being the learning stage implemented on-line using a modified version of the back-propagation through time algorithm Elman (1990); Rumelhart et al. (1986).

All this analysis takes us to the formulation of a second principle of cognitive system construction:

> **Principle 2:** ***Model isomorphism*** — *An embodied, situated, cognitive system is as good as its internalised models are.*

Model quality is measured in terms of some definable isomorphism with the modelled system as established by the modelling relation. Let's see why all this disgression on model learning and quality is relevant for the consciousness endeavour.

3 Reactive vs Anticipatory Control

Many control mechanisms follow the well known error-feedback paradigm. This control structure is so simple and robust that almost all control loops are based on this approach. The strategy is simple and extremely effective Wiener (1961): measure the difference between what we want and what we have and make corrections based on this difference (see Figure 2).

These controllers are very effective but have a serious drawback: they are always *behind the plant*, *i.e.* they cannot make the plant strictly follow a reference signal without a delay (except for special plants in special circumstances). These controllers just act as reaction to plant output diverting from what is desired (errors); so they will wait to act until output error is significant.

In order to have the plant in a certain state at a defined time, we need other, more powerful approaches that can anticipate error and prevent it. Due to the inherent dynamics of the plant, the only possibility of acting to make it reach a final state $\mathbf{s}_f$ at t_f from an intial state $\mathbf{s}_i$ at t_i is to act at t_a before t_f.

This kind of control is anticipatory in this strict sense of ($t_a < t_f$)[7]. The determination of the action cannot come from the final state (as with classical error feedback) because of anticipation and we need an estimate of this state $\hat{\mathbf{s}}_f$ at time t_a.

These two alternative approaches were described by Conant Conant (1969) as *error-controlled regulation* and *cause-controlled regulation*. The advange of this second approach is that in certain conditions, it is often possible for the regulation to be completely succesful at maintaining the proper outcome. Needless to say is that due to the non-identity between model and reality, this last one may depart from what the model says. In these conditions only error-driven control will be able to eliminate the error. This is the reason why, in real industrial practice, model-based controllers are implemented as mixed model-driven and error-driven controllers.

The previous analysis take us into the formulation of another principle:

> **Principle 3:** ***Anticipatory behavior*** *— Except in degenerate cases, maximal timely performance can only be achieved using predictive models.*

These predictive models can be explicit or implicit in the proper machinery of the action generation mechanism Camacho and Bordons (2007). Obviously the degree to which a particular part of reality can be included in a model will depend on the possibility of establishing the adequate mappings from/to reality to/from model and the isomorphims between entailments at the model level and at the reality level (according to a particular model exploitation policy). The problems associted to inferred model quality have been widely studied in relation with properties of statistical modelling, where we seek a good model to approximate the effects or factors supported by the empirical data in the recognition that the model cannot fully capture reality Burnham and Anderson (2004). This is also the world of systems identification but in this case, the target model typically belongs to a very reduced and precise class of models Ljung (1998); Nelles (2000).

4 Integrated Cognitive Control

Reactive and anticipatory control are the core building blocks of complex controllers. Reactive controllers are simpler and more easily tuneable. These are the

[7] This could be seen as acausal because the cause of the action —final cause in aristotelian sense— is the final state $\mathbf{s}_f$, that is a future state.

reasons for being the most used both in biological systems (they are easily evolvable) and technical systems (they are easier to design and implement).

Complex controllers organise control loops in hierarchical/heterarchical arrangements that span several dimensions: temporal, knowledge, abstraction, function, paradigm, *etc.* Sanz (1990). These organisational aspects lead to the functional differences offered by the different achitectures (in the line of Dennet's skinnerian/popperian/gregorian creatures Dennett (1996)).

In the performance of any task by an intelligent agent there are three aspects of relevance: the task itself, the agent performing the task and the environment where the task is being performed Sanz et al. (2000). In the case of natural systems the separation between task and agent is not easy to do, but in the case of technical systems this separation is clearer if we analyse them from the perspective of artificiality Simon (1981). Artificial systems are made on purpose and the task always comes from oustide of them: the owner.

The knowledge content of the models in highly autonomous cognitive controllers should include the three aspects: system, task and environment. Depending on the situation in the control hierarchy, models may refer to particular subsets of these aspects (*e.g.* models used in intelligent sensors do address only a limited part of the system environment; just environmental factors surrounding the sensor).

System cohesion may be threatened in evolutionary terms and its preservation becoments a critical integrational requirement. The problem of model coherence across the different subsystems in a complex control hierarchy is a critical aspect that is gaining increased relevance due to the new component-based strategies for system construction. In the case of biological systems and unified engineering artificial systems the core ontology —whether explicit or assumed— used in the construction of the different elements is the same. But, in systems agregated from components coming from different fabrication processes, ontology mismatches produce undesirable emergent phenomena that lead to faults and even loss of system viability. This is clear in biological systems (*e.g.* immunity-related phenomena) but is just becoming clear in complex technical systems during recent times Horn (2001).

This analysis lead us to formulate an additional principle of complex cognitive systems:

> **Principle 4:** ***Unified cognitive action generation*** — *Generating action based on an unified model of task, environment and self is the way for performance maximisation.*

Modeling the task is, in general, the easiest part[8]. This has been one of the traditional focus points of classic AI (obvioulsy with the associated problem solving methods).

[8] But representing the task in the internalised model can be extremely complex when task specification comes in natural language.

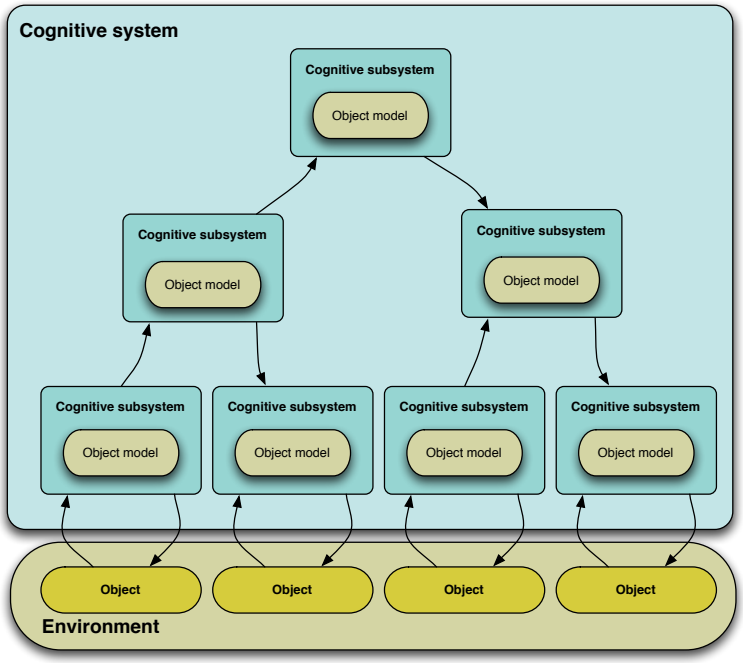

Figure 3: Complex cognitive systems in integrated control architectures need to exploit models in the performance of tasks at different levels of abstraction; from the immediate reaction to environment changes to the strategic decision making relevant for the long term performance of the system.

Modeling the environment in control systems has been generally done up to the extent of addressing the interference it produces in the performance of the task. This can be as simple as statistically modeling an interfering disturbance in SISO controllers (See Figure 2) or as complex as simultaneous localisation and mapping in autonomous mobile robotics.

The question of modeling the system is trickier and will be the main focus of the rest of this paper. Let's say that in conventional analyses of control systems these *realisational* aspects are comonly neglected or reduced to considerations concerning design constraints derived from implementation limitations. The issue of embedding system models —*i.e.* of the system knowing about its own body— has been raised in many contexts but got wider audience in relation with robotics embodiment considerations Chrisley and Ziemke (2002).

5 The Perceiving Agent

As deeply analised by López López (2007) there are strong differences between sensing and perceiving, related to the expectation and model-driveness of this last one.

The perceptual process is structured as a potentially complex pipeline of two classes of processes that we could describe as sensor-driven and model-driven. The perceptual pipeline can affect the perceiving system in two ways: implicitly, through changes in operational states of other subsystems; and explicitly through cognitive integration of what has been perceived into integrated representations.

This unified understanding of perception as a model-driven process López et al. (2007) leads to the introduction of a new principle:

> **Principle 5:** ***Model-driven perception*** — *Perception is the continuous update of the integrated models used by the agent in a model-based cognitive control architecture by means of real-time sensorial information.*

This principle implies that the result of perception is not a scattered series of independent percepts, but these percepts fully incoprorated into an integrated model. This means that it is possible to sense without actually perceiving; *e.g.* if the cognitive —*i.e. model-driven*— sensory processing fails in the integration.

To be integrable, the percept must follow some rules that are captured both in the mechanics of cognitive perception and in the set of referents used in the perception process. The mechanics typically will form part of the permanent *structure* of the agent while some of the referents may be part of its *program* (see Klir (1969) for details on the duality structure/program).

Even more, the perception mechanism is not restricted to process information coming from the environment of the perceiving system but can exploit also infor-

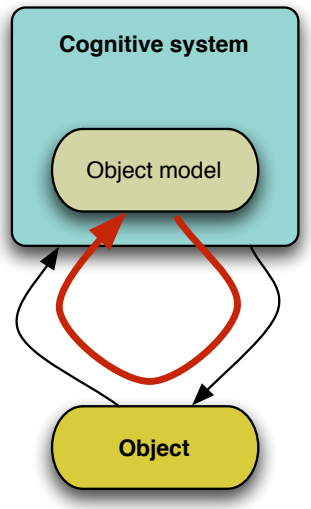

Figure 4: System perception implies the continuous update of the models that the system is employing in the generation of behavior.

mation coming from the inside of the system. Here authors will typically talk about two clasess of preception, *propioception* —the sensing of the body— and *metaperception* —the sensing of the mind— but both are, *senso stricto*, the same class of perceptual processes. This unified perspective implies that for explicit perception to happen in the inner environment, there must be a model where percepts are to be integrated. These models obviously constitute the very core of *self*.

6 The Emergence of Self

As was mentioned before, maintaining system cohesion becomes a critical challenge in the evolutionary trajectory of a cognitive system. From this perspective, the analysis proceeds in a similar way: if model-based behaviour gives adaptive value to a system interacting with an object, it will give also value when the object modelled is the system itself. This gives rise to metacognitition in the form of metacontrol loops that will improve operation of the system overall.

Apart of the many efforts in the analysis of reflective mental processes in biological systems that we are not going to analise in detail here[9], there are also many research threads that are leading to systematically addressing the question of embedding self-models in technical systems. Some of them are:

[9]See for example Gallagher and Shear (2000) for philosophical, cognitive science perpesctives and Kircher and David (2003) for neuroscience and psychology ones.

- System fault-tolerance has been addressed by means of replication of components to avoid single critical failure points; but the determination of faulty states to trigger re-configuration has been a problem of increasing importance in correlation with increased system complexity. Fault detection and isolation methods have developed sophisticated model-based reasoning mechanics to do these tasks. The models used, however, are specifically tailored to the task —a common problem elsewhere.

- Cognitive systems research has put consciousness back into the agenda after many years of ostracism G.A.Mandler (1975) and hence it is addressing the question of computer-based model building of this phenomenon.

- Information systems security —regarding human intrusion and the several varieties of exo-code— has become concerned about the question of self/nonself distinction in ICT systems Kennedy (2003).

- Information systems exploitation is fighting the scalability problem in maintenance tasks trying to mimic the scalable organisation of biological systems Horn (2001)

In our context, control systems, our main concern is not of human mimicking or reduction of cost of ownership. The question is more immediate and basic: system robustness.

There are many technical systems that we depend upon: from the electrical networks, to the embodied pacemakers or ESPs in our cars. Dependability is a critical issue that is being hampered by the increased complexity of individual systems and from emergent phenomena in interconnected ones.

The justifiable quest for methods for managing reasoning about selves in this context is driven by the desire of moving responsibility for system robustness from the human engineering and operation team to the system itself. This is also the rationale behind the autonomic computing movement but in our case the problem is much harder as the bodies of our machines are deeply embedded in the physics of the world.

But the rationale for having self models is even deeper than that: if model-based control overpasses in capabilities to those of error-based control, the strategy to follow in the global governing of a concrete embedded system is not just recognising departure from setpoints but anticipating the behavior emerging from the interaction of the system with it surrounding reality.

Hence the step from control systems that just exploit models of the object, to control systems that exploit models of the pair system + object is a necessary one in the ladder of increased performance and robustness. This step is also observable in biological systems and while there are still loads of unsolved issues around, the

core role that "self" plays in the generation of sophisticated behavior is undeniable. Indeed, part of the importance of self-consciousness is related to distinguishing oneself from the emvironment in this class of models (*e.g.* for action/agency attribution in critical, bootstrapping learning processes).

7 Stepping to Awareness and Consciousness

Our strategy for consciousness research is not following Alexander's approach of axiomatising consciousness (*i.e.* searching for a complex predicate to be accepted by the community) but of analysing and formalising the core issues and mechanisms involved (and addressed in the principles on cognitive control exposed so far). The reason is simple; while the axiomatisation process is important for clarifying the issues, it may give rise to a receding horizon phenomenon similar to what happened to AI in the eighties. This not happening, both approaches should lead to the very same end.

7.1 Defining awareness

From the analysis of integrated cognitive controllers given in the previous sections we can make a try into the formalisation of some consciousness aspects. We will make a distinction between awareness and consciousness, reserving the C-word for systems self-awareness.

> **Principle 6:** ***System awareness*** *— A system is aware if it is continuously perceiving and generating meaning from the countinuously updated models.*

The term *meaning* was introduced in this principle to define awareness and this looks-like eluding the core definitional problem. However, the word *meaning* implies that the main difference between perception and awareness is the addition to the perceptual mechanics of a certain *value system* in the global system process. So we can say that awareness implies the perception of value to the system from its sensory flow.

The updated integrated model produced by perception is *evaluated* in terms of a value system not only in the present state of affairs but in the potential —future and past[10]— consecuences derived from this state of affairs. Awareness implies the partitioning of predicted futures and postdicted pasts by a value function. This partitioning we call *meaning of the update to the model*. In this context of interpretation of the term meaning, we conclude that only pieces of information that are model-integrable can have meaning, because for others, we cannot compute futures nor pasts, less their value.

[10] Restorpective prophecy in the words of T.H. Huxley Huxley (1880).

System perception implies the continuous update of the models that the system is employing in the generation of behavior; but this continuous update is not just keeping in mind an updated picture of the status of part of the environment —like a photograph— but continuously restructuring and retuning the dynamical model of the object used in the action generation process.

System awareness requires the additional steps of automatically predict and evaluate. While many researchers claim for a —necessary— sensory-motor profile of awareness and consciousness, action is not necessary for the definition of awareness; but obviously when the models are used for action selection and built by a process of sensory-motor interaction, action becomes critical for the awareness architecture; but models can be built using other methods (see Section 2.4) and this will be more manifest in artificial systems.

7.2 Defining consciousness

When the target of the awareness mechanism is the aware system itself, consciousness happens:

> **Principle 7:** ***System self-awareness/consciousness*** *— A system is conscious if it is continuously generating meanings from continously updated self-models in a model-based cognitive control architecture.*

System self-awareness —consciousness— just implies that the continuous model update include the updating of submodels about the system itself that are being evaluated. The models of the supersystem —system+object— are used in the model-based generation of system behavior. So the process of behavior generation is explicitly represented in the mind of the behaving agent as driven by a value system. In this sense the interpretation of consciousness that we propose here depart from higher-order theories of consciousness Rosenthal (ming); Kriegel and Williford (2006) in the fact that self-awareness is not just higher order perception. Meaning generation is lacking in this last one.

Another question of extreme relevance is the maximally deep integration of the model and metamodel. As Kriegel Kriegel (2006) argues, higher-order monitoring theory makes the monitoring state and the monitored state logically independent with a mere contingent connection. We are more in the line of Kriegel same-order monitoring theory that argues for a core non-contingent relation between the monitoring state and the monitored state.

One big difference between being aware and being conscious cames from the capability of action attribution to the system itself thanks to the capability of making a distinction between self and the rest of the world[11]. This implies that a conscious

[11]Obviously, even while we argued for awareness/consciousness as a purely input, perceptual process, these associations to action processes links consciousness with action generation and even

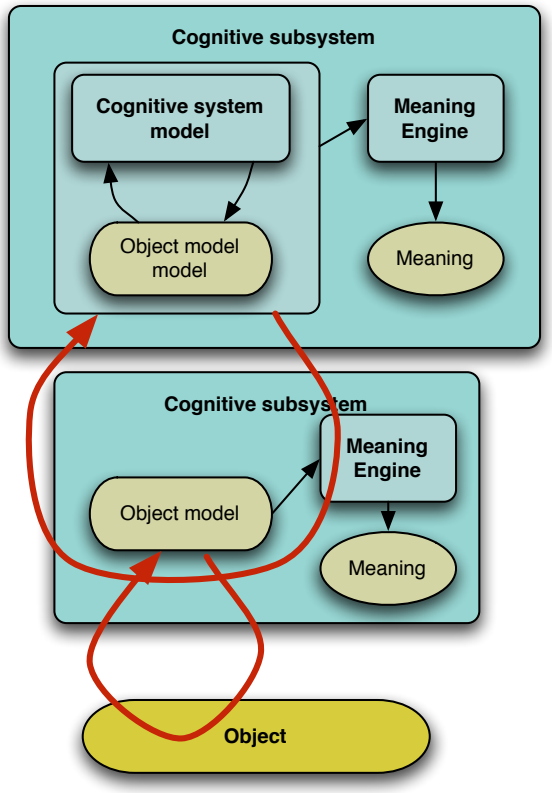

Figure 5: System self-awareness —consciousness— implies the continuous update and meaning generation of a behavior-supporting model that includes a submodel of the very agent generating the behavior (apart of a model of the model of the object).

agent can effectively understand —determine the meaning— the effect of its own actions (computing the differential value derived from self-generated actions, *i.e.* how its own actions change the future).

Even more, conscious agents can be made responsible and react to past actions by means of retrospectively computing values. So, a conscions agent will be able to understand what has been its role in reaching the actual state of affairs.

This appreciation of the coupling of consciousness to value systems is also being done in biological studies of consciousness. Let's quote Gerald Edelman Edelman (2006) at this point:

> *Consciousness appeared in vertebrate evolution when reentrant connections in the thalamocortical system arose to link anterior memory systems deal-*

with system's ethics.

ing with value to the more posterior cortical systems devoted to perception. The result was an enormous increase in discriminatory power resulting from myryad integrations among the reentrant circuits comprising this dynamic core.

7.3 Addressing Aleksander's five axioms

Being conscious (continuously updating and computing meaning from world-self-models in a model-based cognitive control loop) has some implications that we can analyse in the context of Aleksander's five axioms for consciousness Aleksander and Dunmall (2003) regarding a system A in a world S:

Axiom 1 — Depiction: *"A has perceptual states that depict parts of S."* The model of S that A has is continuously updated and contains a deep structure that makes it a depiction of S.

Axiom 2 — Imagination: *"A has internal imaginational states that recall parts of S or fabricate S-like sensations. "* Models are used in predictive-causal exploitation engines that may be used to predict, postdict or perform counterfactual analyses of what-if situations.

Axiom 3 — Attention: *"A is capable of selecting which parts of S to depict or what to imagine. "* Effective use of scarce reasoning and metareasoning resources would imply that the system has to select part of the whole specturm of awareness flow for further analysis.

Axiom 4 — Planning: *"A has means of control over imaginational state sequences to plan actions. "* The stucture of the cognitive system based on principles 1-6 implements a metacontrol schema with full anticipatory capabilities for future action generation.

Axiom 5 — Emotion: *"A has additional affective states that evaluate planned actions and determine the ensuing action."* Utility functions at metalevels are used to decide between alternative control strategies.

8 Conclusions

This strategy of control based on self models can be applied recursively and have metacontrol systems based on recursively applying the model-based machine structure. This is obviosly related to Sloman and Chrisley vision on virtual machines and consciousness Sloman and Chrisley (2003). In our model, however, passing from the first level to the next ones, something interesting happens. As Gell-Mann suggests Gell-Mann (2001):

> "At successive levels, it is the availability of similar mathematical descriptions from related problems that makes the next step appear with simplicity and elegance."

Successive levels may have self-similar structures —even when the concrete realisations may appear entirely different— and hence it is possible to think on closing the progressive recursiveness if the model of the metalevel is able to capture its own core structure; *i.e.* a particular control level is able to reason about itself as a metalevel controller (this is what is proposed by Krieger Kriegel and Williford (2006)). The system then becomes cognitively closed and the metareasoning can progress ad infinitum —up to the limits of resources — without the need on new implementational substrates.

As final conclusions of this paper, let's comment on some of the proposed focus points sugested by Taylor and mentioned in Section 2.1:

1. *General principles for cognitive systems*. Some proposals for principles are introduced in this paper. All of them centered around the issue of model-based action generation up to the level of reflective, model-based action.

2. *The pros and cons of embodiment for cognitive systems*. From the reasons shown above, cognitive systems are necessarily situated and are necessarily embodied to the extent of the action generation resources and the self-awareness capabilities necessary for unified cognitive action generation (which is necessary for preserving system cohesion).

3. *The desirability or otherwise of guidance from the brain*. The brain demonstrates the effectivity of the principles shown above. This does not imply that building direct copies of the brain is the best approach but extracting structural patterns that can be analysed formally.

4. *Specific cognitive system designs and their powers*. We have summarily analysed three cognitive systems designs —cognitive, metacognitive and reflective— and have argued for the possibility of closing the list of progressively meta structures at this last one.

5. *The best future pathways for development of cognitive systems*. Explore the issue of cognitive architectures based on hierarchical unified modeling of the system/object structures.

9 Acknowledgements

We acknowledge the support of the European Commission through the grant *ICEA: Integrating Cognition, Emotion and Autonomy* and of the Spanish Ministry of Educa-

tion and Science through grant *C3: Conscious Cognitive Control*. Details on both can be found at our website: `www.aslab.org`.

We also want to thank the editors of the special issue of Neural Networks where this article was published —John Taylor, Walter Freeman and Axel Cleeremans— for the invitation to contribute to it.

References

Aleksander, I. and Dunmall, B. (2003). Axioms and tests for the presence of minimal consciousness in agents. *Journal of Consciousness Studies*, 10(4-5):7–18.

Blackmore, S. J. (1999). *The Meme Machine*. Oxford University Press.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Burnham, K. P. and Anderson, D. R. (2004). *Model Selection and Multimodel Inference, A Practical Information-Theoretic Approach*. Springer, New York.

Camacho, E. F. and Bordons, C. (2007). *Model Predictive Control*. Springer, second edition.

Cellier, F. E. (1991). *Continuous System Modeling*. Springer-Verlag, New York.

Chrisley, R. and Ziemke, T. (2002). Embodiment. In *Encyclopedia of Cognitive Science*, pages 1102–1108. Macmillan Publishers.

Conant, R. C. (1969). The information transfer required in regulatory processes. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):334–338.

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97.

Craik, K. J. (1943). *The Nature of Explanation*. Cambridge University Press.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Dennet, D. C. (1991). *Consciousness Explained*. Little, Brown & Company.

Dennett, D. C. (1996). *Kinds of Minds: Toward an Understanding of Consciousness*. Science Masters. Basic Books.

Edelman, G. M. (2006). *Second Nature*. Yale University Press, New Haven.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Gallagher, S. and Shear, J., editors (2000). *Models of the Self*. Imprint Academic, Exeter, UK.

G.A.Mandler (1975). Consciousness: Respectable, useful, and probably necessary. In Solso, R., editor, *Information Processing and COgnition: The Loyola Symposium*. Erlbaum, Hillsdale, NJ.

Gell-Mann, M. (2001). Consciousness, reduction, and emergence. some remarks. *Annals of the New York Academy of Sciences*, 929(1):41–49.

Gentner, D. and Stevens, A. L., editors (1983). *Mental models*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, (23):121–123.

Grush, R. (1995). *Emulation and Cognition*. PhD thesis, UC San Diego.

Henriques, J., Gil, P., Dourado, A., and H.Duarte-Ramos (2002). Nonlinear multivariable predictive control: Neural versus first principle modelling approach. In Hamza, M. H., editor, *In Proceedings of Control and Applications, CA 2002*, Cancun, Mexico.

Horn, P. (2001). Autonomic computing: Ibm perspective on the state of information technology. IBM Research.

Huxley, T. H. (1880). On the method of zadig. retrospective prophecy as a function of science. In *Collected Essays IV: Volume IV, Science and Hebrew Tradition*. Project Gutemberg.

Johnson, M. (1987). *The body in the mind*. University of Chicago Press, Chicago.

Kennedy, C. M. (2003). *Distributed Reflective Architectures for Anomaly Detection and Autonomous Recovery*. PhD thesis, University of Birmingham.

Kircher, T. and David, A., editors (2003). *The Self in Neuroscience and Psychiatry*. Cambridge University Press.

Klir, G. C. (1969). *An approach to General Systems Theory*. Litton Educational Publishing, Inc.

Klir, G. J. (2001). *Facets of Systems Science*. Kluwer Academic/Plenum Publishers, New York, second edition.

Kriegel, U. (2006). The same-order monitoring theory of consciousness. In Kriegel, U. and Williford, K., editors, *Self-Representational Approaches to Consciousness*, pages 143–170. MIT Press.

Kriegel, U. and Williford, K., editors (2006). *Self-Representational Approaches to Consciousness*. MIT Press.

Libet, B., Wright, E. J., Feinstein, B., and Pearl, D. (1982). Readiness potentials preceding unrestricted "spontaneous" vs pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology*, 54:322–335.

Ljung, L. (1998). *System Identification: Theory for the User*. Prentice Hall PTR, 2nd edition.

López, I. (2007). *A Framework for Perception in Autonomous Systems*. PhD thesis, Departamento de Automática, Universidad Politécnica de Madrid.

López, I., Sanz, R., and Bermejo, J. (2007). A unified framework for perception in autonomous systems. In *Proceedings of the Cognitive Science Conference 2007*, Nashville, USA.

Nelles, O. (2000). *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer.

Pockett, S. (2004). Does consciousness cause behaviour? *Journal of Consciousness Studies*, 11(2):23–40.

Rosen, R. (1985). *Anticipatory Systems*. Pergamon Press.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. Columbia University Press.

Rosen, R. (1993). On models and modeling. *Applied Mathematics and Computation*, 2-3(56):359–372.

Rosenthal, D. M. (Forthcoming). Higher-order theories of consciousness. In McLaughlin, B. and Beckermann, A., editors, *Oxford Handbook on the Philosophy of Mind*. Clarendon Press, Oxford.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pages 318–362. MIT Press, Cambridge, MA, USA.

Sanz, R. (1990). *Arquitectura de Control Inteligente de Procesos*. PhD thesis, Universidad Politécnica de Madrid.

Sanz, R., López, I., Rodríguez, M., and Hernández, C. (2007). 2007 special issue: Principles for consciousness in integrated cognitive control. *Neural Netw.*, 20(9):938–946.

Sanz, R., Matía, F., and Galán, S. (2000). Fridges, elephants and the meaning of autonomy and intelligence. In *IEEE International Symposium on Intelligent Control, ISIC'2000*, Patras, Greece.

Simon, H. A. (1981). *The Sciences of the Artificial*. MIT Press, Cambridege, MA, second edition.

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciosness Studies*, 10(4-5):133–172.

Snow, C. (1969). *The Two Cultures and a Second Look*. Cambridge University Press.

Varela, F. G. (1971). Self-consciousness: adaptation or epiphenomenon? *Studium Generale (Berl)*, 24(4):426–439.

Wiener, N. E. (1961). *Cybernetics, or Control and Communication in the Animal and the Machine*. MIT Press, second edition.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics*, 13(1).

Wolpert, D., Ghahramani, Z., and Jordan, M. (1995). An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882.

Zeigler, B. P., Kim, T. G., and Praehofer, H. (2000). *Theory of Modeling and Simulation*. Academic Press, 2nd edition.