Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 938–946
Neural Networks
www.elsevier.com/locate/neunet

2007 Special Issue

Principles for consciousness in integrated cognitive control

Ricardo Sanz*, Ignacio López, Manuel Rodríguez, Carlos Hernández

Autonomous Systems Laboratory, Universidad Politécnica de Madrid, 28006 Madrid, Spain

Abstract

In this paper we will argue that given certain conditions for the evolution of biological controllers, they 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.

Keywords: Consciousness; Control architecture; Modelling; Model-based control; Machine consciousness

1. A bit of context

In this our excursion into the scientific world 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 Dennet (1991), Libet, Wright, Feinstein, and Pearl (1982), Pockett (2004) and Varela (1971)). Hence, we may wonder why engineers should be interested in such a phenomenon that is not yet fully understood and somehow not even fully accepted.

In this paper we will argue for a precise interpretation of consciousness – based on controller mechanics – that renders it not only not epiphenomenal but also 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 for 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 fully 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 necessarily appearing trait of sophisticated machines. This is where we are: identifying the core mechanics and application constraints for the realisation of consciousness capabilities in next generation technical systems. This will imply, necessarily, the sound characterisation of the expected benefits from making a machine conscious.

2. The modelling of the 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 & 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;

* Corresponding author.
E-mail address: ricardo.sanz@upm.es (R. Sanz).
URL: http://www.aslab.org (R. Sanz).

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

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.09.012

– Specific cognitive system designs and their powers;
– Embodied cognitive systems in robot platforms and demonstrations;
– The best future pathways for the development of cognitive systems.

Proposing some cognitive principles up to the level of consciousness will be the objective of this paper. Let us 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 knowledge 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 modelling 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 explicitly 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 & Stevens, 1983) but this has just had an aura of metaphorical argumentation (Johnson, 1987) because of the lack of formalisation of the concept of model and 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 (1995) or the model-based sensory–motor integration theory of Wolpert, Ghahramani, and Jordan (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 metaphor 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] The truth of a piece of information included in a model is not just its fitness into the model – e.g. a perspective held by social constructivists – but in terms of the establishment of isomorphims between the model and the modelled.

2.2. On models

This definition of cognition as model-based behavior may sound too strict to be of general applicability; in particular it seems not to fit simple cognitive processes (e.g. it seems that we can have a stimulus input without having a model of it). However, if we carefully analyse 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 successful.

These information structures may be explicit and directly identifiable in their isomorphisms or may be extremely 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 surrounding 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 & Draper, 1987). It is this model usefulness which gives adaptive value to cognition as demonstrated by Conant and Ashby (1970).

There are plenty of references on modelling theory, mostly centered in the domain of simulation (Cellier, 1991; Zeigler, Kim, & Praehofer, 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, 1991, 1993).

This last cited one 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 system 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 surrounding reality.

We must bear in mind, however, that models may vary widely in terms of purpose, detail, completeness, 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 & Ashby, 1970) – every good regulator must contain a model of the system it is