

What Does Consciousness Bring to CTS?

Daniel Dubois⁽¹⁾, Pierre Poirier⁽²⁾, Roger Nkambou⁽³⁾

University of Quebec at Montreal

PO Box 8888, Centre-Ville Station, Montreal (Quebec) Canada, H3C 3P8

(1), (3) GDAC Laboratory, Department of Computer Science

(2) Department of Philosophy

dubois.daniel@uqam.ca, poirier.pierre@uqam.ca, nkambou.roger@uqam.ca

Abstract

Striving in the real world is more and more what artificial agents are required to do, and it is not a simple task. Interacting with humans in general, and with students in specific, requires dealing with a great number of information sources, and high volumes of data, in real-time, to adapt to the environment (and other agents). «Consciousness» mechanisms can help and sustain an apt artificial tutor, allowing it to consider various sources of information in diagnosing and guiding learners. We show in the present paper how they effectively support these processes in the specific context of astronauts training on the manipulation of the Space Station Robotic Manipulation System, Canadarm2.

Introduction

¹Many cognitive models, such as CS/SAS (Norman and Shallice 1980; Norman and Shallice, 1986; Cooper and Shallice, 2001) and TLA (Gat, 1998), propose a three-layer architecture. Although these systems seem able to reproduce some experimental phenomenon observed by psychology experiments about decision and behavior, and may include deliberative capabilities (Glasspool, 2000), they do not attempt to explicitly model the whole cognitive human architecture. Memory systems are not explicitly addressed and, although attention (in some form) has a role in them, consciousness does not. ACT-R (Anderson, 2003) also offers the upper-goals and motor layers, possesses deliberative capabilities, and is very much concerned with memory. However, emotions are absent; if they were to be included in the reasoning phase, then the rules would crank up a level of complexity. ACT-R does not either consider the existence of consciousness, at least not explicitly. None of these architectures mention what can be claimed as the finest adaptation means offered by Mother Nature: consciousness (Sloman, 2003), and they do not either take into account feelings and emotions. We believe a holistic theory of the mind ought to incorporate a role for consciousness and for feelings, and we show how they may contribute to decisions in CTS, a "conscious" tutoring agent.

The paper is organized as follows: first, we clarify the concept of consciousness as we intend to use it. Then we present Baars' *Global Workspace* theory after which our consciousness model is built, and we go on to present the "conscious" tutor's architecture derived from it. In a fourth section, we illustrate how a diagnosis process unfolds through the cooperative work of the various components of the architecture, thanks to the "consciousness" mechanisms.

What Consciousness Is About

Consciousness is a confusing concept (Minsky, 1998). It covers different ideas and phenomena (just as do intelligence, learning, memory, and intuition). Aside from more popular notions, Block's individuation of various forms of consciousness (Block, 1995) clarifies that it includes the internal mechanisms allowing us to represent and make the content of the present experience available to the rest of our internal, unconscious processes. This grants access to otherwise unreachable resources. He calls this phenomenon *access consciousness*. Another aspect of the word aims at those faculties that keep us informed about the activities of our senses; he calls it *monitoring consciousness*. Block naturally also mentions consciousness as including the idea of *self-consciousness* (being aware of existing as an entity distinct from the rest of the world) and the much discussed *phenomenal consciousness*.

What do we do with such a bizarre concept in AI? And in Education? Well, even though it is unconventional and quite an undertaking, equipping an agent with "consciousness" mechanisms might grant it with the best adaptation means nature has to offer to face a complex environment (Sloman, 2003). For this enterprise, a well-rounded theory would be a good place to start. Baars provided it in 1988. Our agent proposes a functional implementation of monitoring and access consciousness as described by Baars, and has no pretense to address what Chalmers coined "the *hard problem*" of consciousness (Chalmers, 1995). For the time being, we simply aim at a useful reproduction; accordingly, we keep the word inserted in quotations marks!

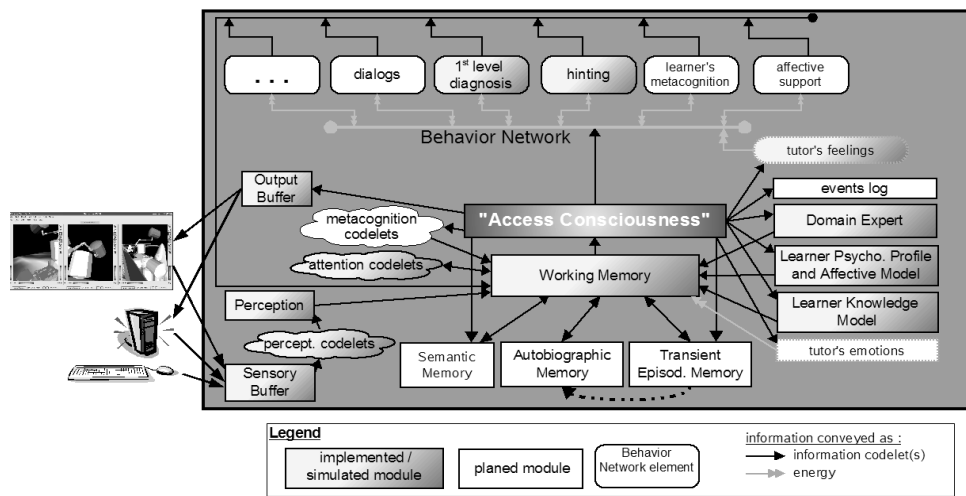


Figure 1 CTS' conceptual architecture. Modules communicate exclusively with Working Memory and receive all their information from it through the work of Attention and Access "consciousness" mechanisms.

Baars' Global Workspace Theory

Baars (1988, 1997) has proposed a theory that unifies many previous efforts in describing and modeling the human mind and human consciousness. In his view, consciousness plays fundamental roles in nine functions, among which we find: Adaptation and learning, Contextualizing, Prioritizing and access control, Recruitment, Decision-making and Self-monitoring. All these functions, and all brain operations, are carried by a multitude of globally distributed, unconscious specialized processors (which we implement computationally as *codelets*; to be explained in the section about the Behavior Network and codelets). When one of them cannot complete its operation, it tries to make this situation known to all other processors, in other words have the whole system become conscious of the situation. Processes that recognize the fact and know what to do about it, or how to take over from this step, grab a copy of the information and process it; there is no need for any central coordination mechanism). What part of the system will effectively respond is unconsciously shaped (influenced) by the context: current goal and plan, expectations, current mood and emotion, personal preferences, complementary information brought back by other systems such as memories or emotional systems, or processes currently in the forefront. As a small example of the role of context, if we hear the word "set", we won't interpret it and react to it the same way if we are watching a tennis match, attending a mathematical lecture, or being asked to set the table (Baars, 1988).

A situation brought to consciousness for an explicit, collective processing, is described by coalitions of processors presenting various aspects of it. Many such coalitions may compete to gain access to this global workspace of limited capacity.

Let us now see how this theory can be implemented and used to build a tutoring agent.

Our Implementation of Baars' Theory

Description of the Architecture

Our architecture is essentially rooted in Professor Franklin's IDA's architecture (Franklin, 2005) and owes much to it, but brings some domain specific extensions (such as learner modeling) and some modifications to the implementation. As does IDA (and LIDA, its "learning" offspring), its conceptual architecture (**Figure 1**) covers every major aspect of cognition, with many functional parallels to the physiology of the brain (Franklin *et al.*, 2005). Not readily visible on the diagram are some processes such as action selection, learning, deliberation. At the center of all this are Attention, Access "Consciousness", and the computational reproduction of Baars' mind processes (or neuronal groups): the various types of *codelets*. Here's a description of main aspects of the architecture.

Senses and Perception. The Sensory Buffer serves as an inwards interface to any external actor. Every dynamic aspect of the "environment" appear in the messages received from the Simulator: Canadarm2 configuration (rotation angle of every joint), position of the payload, camera selected on each of the three monitors along with its dynamic attributes (zoom, pitch and yaw angles), etc. If the event was not manipulation related, other types of information are supplied, such as exercise type and specifications. The perceptual codelets scan the buffer and activate nodes in the Perceptual Network (PN); those nodes receive data and give it semantic meaning through the hierarchical organization of the network (resulting in "concepts" the agent can recognize: "Canadarm2 manipulation", "user answer", etc.). They also grant it importance on a semantic basis. A grammar has been developed to implement the communication between the Simulator and the Sensory Buffer.

Learner Model. An extension to IDA, the Learner Model is spread over three modules. In its static part, the Learner Profile (LP) contains psychological information, including learner's learning style. Its dynamic part, Learner's Affective State (LAS), tracks learner's mood and emotional state. Learner Knowledge Model (LKM) holds facts, infers knowledge and trends, and computes statistics. They all volunteer information when they deem appropriate, eventually priming some "feeling(s)" in the agent (to be explained in the subsection about the "personality" of the agent). The LKM is the main mechanism in establishing the causes of the learner's difficulties. Pending realization, it will very likely be implemented as bayesian networks. Their factual knowledge nodes will record evidences found in the Access "Consciousness" broadcasts.

The Behavior Network (BN) and the Codelets. Based on an idea from Maes (Maes, 1989) and modified by Negatu and Franklin (Negatu and Franklin, 2002), the Behavior Network holds the repertoire of the agent's know-how in the form of streams of behavior nodes, and offers a way to decide on which should activate. Each behavior specifies its necessary preconditions and indicates the effects it should have on the environment. Nodes accumulate the energy that comes from the agent's "feelings and desires" (some special high-level goal nodes that are meant to play the role of "feelings" and "desires" in planning and deciding) until they are elected for action. That arrangement accomplishes high-level planning and also creates a stable global behaviour for the agent. "Desires" aim here at representing professional and personal goals of the agent, that is, high-level goals that CTS pursues not simply in reaction to the environment. Some of these may be to "Keep learner affective state high", or "Sleep" (so that the agent may find the time to do some heavy analysis not possible on-line).

The BN serves as the coordinator for the agent's actions, and generally counts on other functionalities' reactions to render a service. As an example, when all the conditions are met, the Diagnosis stream declares the need for the probable cause, and LKM will very likely respond to this request "heard" in the broadcast.

Negatu and Franklin also modified Maes' model so that each behavior is realized by a collection of *codelets*. Taken from the copycat architecture (Hofstadter and Mitchell, 1994), codelets are simple agents implemented on a few tens of line of code, limited in their abilities (mostly pattern recognizers with some action capabilities) but highly efficient. They quite nicely correspond to Baars' simple processes of the mind. These do not appear explicitly on the diagram yet are essential to the architecture. They accomplish a major part of the operations in the agent, they render effective many functions, and they connect all of the modules to Access "Consciousness". For instance, information codelets carry information from one place to another, and expectation codelets make sure expected results are met. Codelets inherit the energy value of the information they carry (and/or receive energy from the BN node they implement). A collision risk, for in-

stance, bears a high informational value (established by the designers); it gives the coalition of information codelets describing it better chances of being selected for broadcasting.

The Personality of the Agent. "Feelings" and "Desires", in our architecture, are the basic mechanisms forming the agent's personality. Emotions are only in the design phase. "Feelings" and "Desires" are the motivational mechanisms that feed the Behavior Network with activation, and so orient action selection in line with the agent's high-level goals. They contain CTS' *attitudes* towards various situations (its predispositions to react in such and such way), so its *personality traits*. Emotions are also expected to intervene in the unconscious decision process taking place in BN, but also during memorization, learning and deliberations. Examples of specific situations that stimulate some "Feeling" of the need for a specific type of action may appear when the LKM signals some important flaw in the learner's knowledge, or may appear after the broadcast of a perceived external situation, such as the possibility of a collision while the astronaut is manipulating Canadarm2.

Attention and Access "Consciousness". Attention selects the most important coalition of codelets in WM, and Access "Consciousness" broadcasts its information, allowing all other systems to become aware of the situation. This mechanism is crucial for the collaboration of the parts, for instance in reaching a diagnosis.

Some Functional Aspects of the Architecture

Considering various aspects from various sources is very natural for our "conscious" architecture: every "module" is made aware of important elements and can contribute information describing aspects of the situation or useful towards reaching a decision. For example, when the tutor (CTS) perceives a lack of action from the learner, it will try to determine its nature and cause. Its Episodic Memory will automatically (try to) recall somewhat similar situations and bring them back into Working Memory (WM) so that they can be analyzed for commonalities; the LKM and the Domain Expert may give hypotheses too. If no explanation shows up, that situation may stimulate CTS' Feeling node for the «Need to interact with the user» to clarify the case (interactive diagnosis).

The internal debating implemented offers great flexibility in behavior adaptation, even more so since the architecture can accommodate any number of "modules" intervening towards finer behaviors. We do not have the necessary room to explain the process of coalition formation and selection, but we can briefly mention that there are two processes involved: coalitions that get enriched by modules grafting information codelets to them, and coalitions that eventually form by Hebbian learning in WM. The former has to do with deliberations and decisions, where attention and access "consciousness" are essential; the latter has to do with learning the environment, but is an attribute of the global architecture, not specifically of consciousness. In either case, they are made possible by Baars' idea of collaboration between the mind's simple processors.

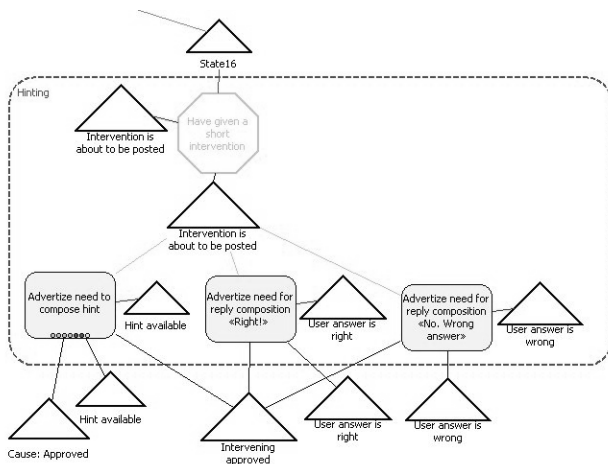


Figure 4 Excerpt from the Behavior Network. The Hinting stream shows three preconditions for its "Compose a hint" sub-stream: Cause approved, Hint available, and Intervening approved.

proved. But it is the Domain Expert that holds hints related to the knowledge about procedures.

The coalition just published remains active in WM and new information may attach to it during the forthcoming deliberation. As long as this coalition is the most important one in WM, it is broadcast repetitively until suggested causes do arrive into WM, or until the Deliberation Arbiter determines that enough time has passed without any change to the coalition. In our scenario, the LKM and LPM modules have hypotheses to offer, after the time needed for their inference process: «Poor mastery of manipulation procedure» and «Distracted».

Only one cause can attach to the original coalition. When confronted with many possible causes (offered by different sources), the Deliberation Arbiter selects the most probable cause. The probability of a cause is obtained by multiplying the cause's value with the confidence on the hypothesis. The "poor mastery" hypothesis is retained here. The Arbiter attaches that cause to the coalition, which adds new activation to it. This association mechanism implements Baars convergence of information phenomenon (Baars, 1997). If this coalition is selected in WM and is broadcast, the new aspects in the information should prompt new reactions in the *audience* (the modules and codelets hidden in the unconscious, as proposed by Baars' Theatre metaphor). Here, the feeling for intervening gets more stimulation from it. Some module could also react and oppose the cause proposed. This would stimulate the module that saw its hypothesis refused to submit a new cause, extending the deliberation process. An opposition could also aim plainly at the idea of intervening. A number of reasons could justify such opposition:

- a module (LPM or LAM) estimates it would be damaging to intervene in the actual state of mind or affective state of the learner;
- there is no cause (or no sufficient cause) for it;

- the support level chosen by the learner does not warrant intervening here.

An opposition to intervening simulates the experience we all have had of planning on intervening (for example, replying something nasty to someone) and just before the words went out of our mouth, refraining from doing so. It reflects parts of William James' ideomotor theory (Baars, 1997).

In this simple scenario, nothing of the sort happens. The "standard" waiting time of five cognitive cycles is respected, during which the coalition is published repetitively with the proposed cause. The cause is not opposed, and neither is the idea of intervening. So, the Arbiter changes the status of both the cause and the proposition to intervene to «Approved». Then, the Arbiter knows it has completed its task. The broadcast that ensues stimulates the «Intervening approved» State in the BN (see **Figure 4**). The proposition of intervening is implicitly sustained by the LPM by not opposing it and by indicating the user's preferred way of interaction: hinting.

An attention codelet keeps note of the number of hints previously given in this intervention and replies with the hint number to request: «Hint to give: 1». With this last information about how to interact with the astronaut, the Domain Expert is able to offer a contribution in the form of a hint appropriate for the situation (based on the problem observed, the actual status of the manipulation, and the previous hint given). A first-level, very general hint is suggested, but its content is not directly shown to the astronaut, but sent into WM.

When published, that content brings the final needed context in the Behavior Network for hinting as an intervention: the activation of the State «Hint available». When this information becomes available, the behavior appearing to the left in the Hinting stream can send away its codelets, requesting that a hint with that content be put into shape and shown in a window on screen. A specialized process will eventually take care that, taking necessary steps to show on the simulator's monitor «Hint: Haven't you forgotten something?».

From this point, CTS will continue with further hinting, progressively more specific and instructional, until the astronaut corrects the situation. Or until something of greater importance appears in Working Memory, gets selected and published, and brings other resources to react and cause a new stream of actions.

Concluding Remarks

Without attention and consciousness mechanisms or their equivalent, an artificial agent processes all that its sensors receive, and it needs a central coordinator to organize the collaboration of the parts. In comparison, CTS filters its inputs at the perceptual level, selects the most important information in Working Memory, and submits it to every resource in the architecture. Only relevant resources react to it (Franklin, 2005).

CTS' architecture allows great flexibility for adapting to the environment, thanks to its attention and consciousness functional mechanisms. They are what allow non-automatic actions to be recruited when automatic processes are not enough to cope with the situation at hand. They are involved in building new plans, in adapting or modifying the existing ones. The deliberation process that consciousness makes possible, among other roles, enables the identification of the most important information in the context; it also enables iterative enrichment of a decision, or the inhibition of an action. Together, these ideas support the belief that consciousness is the most powerful means of adaptation grown through Evolution, although not the only one.

Through a walk-through scenario, we have given an illustration of how access consciousness can be involved in the agent's adaptation. Much more could be said and explained. Other types of consciousness exist in this architecture, at least to some levels, and could have been singled out. But we mostly wanted to illustrate that AI could help to implement a theory of the mind and of consciousness, and that this implementation could benefit AI with the potential for a highly adaptable agent.

Much work remains to be done both on the cognitive/conceptual side and in the implementation side of our prototype. Emotions are still in the design phase, some forms of learning are in the integration phase but memories exist only in the Domain Expert module, and we will need to lean on formalizing many aspects of the architecture. But seeing an embryo of a conscious agent gives our whole team motivation to pursue the efforts.

Acknowledgments. Special thanks to Professor Franklin and to the University of Memphis, who graciously granted us access to their IDA technology. We also sincerely thank the following members of our team that are an integral part of our research effort: Mohamed Gaha, Usef Faghihi, Philippe Fournier-Viger, Patrick Hohmeyer.

References

Anderson, J. R.; Bothell, D.; *et al.* 2004. An integrated theory of the mind. In *Psychological Review* 111 (4): 1036-1060.

Baars, Bernard J. 1988. *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Baars, B. J. 1997. *In the theater of consciousness: The Workspace of the Mind*. New York, NY: Oxford University Press.

Block, N. 1995. On a Confusion about a Function of Consciousness. In *The Behavioral and Brain Sciences* (18).

Cooper, R., and Shallice, T. 2001. Contention Scheduling and the control of routine activities. *Cognitive Neuropsychology*.

Franklin, S. 2005. A "Consciousness" Based Architecture for a Functioning Mind". In *Visions of Mind*. Davis, D. eds. Hershey, PA: IDEA Group, Inc.

Franklin, S.; Baars, B. J.; *et al.* 2005. The Role of Consciousness in Memory. In *Brains, Minds and Media*. vol. 1.

Gat, E. 1998. On three layer architectures. In Kortenkamp, D.; Bonasso, R. P. ; and Murphey, R. eds. *Artificial Intelligence and Mobile Robots*. AAAI Press.

Glasspool, David W. 2000. The integration and control of behaviour: Insights from neuroscience and AI. In *Proceeding of AISB Convention*.

Hofstadter, D. R., and Mitchell, M. 1994. The Copycat Project: A model of mental fluidity and analogy making. In *Advances in connectionist and neural computation theory, vol. 2: logical connections*, Holyoak, K. J., and Barnden, J. A. eds. Norwood N.J.: Ablex.

Maes, P. 1989. How to Do the Right Thing. In *Connection Science Journal* 1 (3): 291-323.

Minsky, M. 1998. Consciousness is a Big Suitcase: A Talk with Marvin Min-sky. In Edge, February 1998. On line. http://www.edge.org/3rd_culture/minsky/minsky_p2.html. Retrieved on November 8, 2006.

Negatu, A. S.. and Franklin, S. 2002. An Action Selection Mechanism for "Conscious" Software Agents. In *Cognitive Science Quarterly* 2: 363-386.

Norman, D. A., and Shallice, T. 1980. *Attention to action: Willed and automatic control of behaviour*. Center for Human Information Processing (Technical Report No. 99). San Diego: University of California.

Norman, D. A., and Shallice, T. 1986. Attention to action: Willed and automatic control of behaviour. Reprinted in revised form in R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.) 1986. *Consciousness and self-regulation*, Vol. 4 (pp. 1-18). New York: Plenum Press.

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