

Achieving Advanced Machine Consciousness through Integrative, Virtually Embodied Artificial General Intelligence

Ben Goertzel

*CEO, Novamente LLC
Rockville, Maryland, USA*

Abstract

In my talk I will deal with two somewhat separate issues: 1) What is consciousness (and what is “advanced machine consciousness”)?, 2) What is my plan for achieving advanced machine consciousness?

1. What is consciousness (and what is advanced machine consciousness)?

“Consciousness” is a subtly polysemous English word; and as with other subtly polysemous words like “intelligence,” “love,” “beauty” and so forth, it wraps up a complex network of interrelated concepts. It would be a mistake to assume that, just because a certain concept-network has been associated with a single word by a certain set of cultures (note that the English word “consciousness” has no precise correlate in e.g. Chinese, Sanskrit or Australian aboriginal languages), it must necessarily have a fundamental philosophical or empirical meaning. Words arise for all sorts of reasons, but largely because of their utility for a certain group in a certain cultural context.

Like many others, I find it useful to distinguish multiple perspectives on the world, including the subjective (first person), intersubjective (second person) and objective (third person) perspectives. What Chalmers and others call “raw consciousness” has to do mainly with first and second person perspectives. “Neural correlates of consciousness” have to do with patterns binding together first-person with third-person phenomena. For instance, in *The Hidden Pattern* (where this multi-perspectival viewpoint is outlined in more detail), the following hypothesis is made: in cases where a network of subjective experiences is correlated with a specific physical system, then subjective experiences that are more experientially intense, generally correspond to patterns in the physical system that have more informational significance (where “informational significance” may be measured mathematically in many ways, e.g. using a formula called the “pattern intensity”). This sort of correlation does not reduce either of the first- or third- person perspectives to the other, but simply relates the two. The language used to perform this interrelationship is that of correlations and patterns.

Often the term “consciousness” is used to denote a particular set of mental structures and dynamics that occur in human minds, in correlation with particularly intense subjective experiences. These structures and dynamics include things like reflection, deliberation, short-term memory, working memory, self, will, and so forth. My view is that these structures and dynamics are best viewed as emergent phenomena that emerge from the complex dynamical system of the human mind-brain. They evolved in humans in large part because they served the goals of the human organism (though with some spandrelization involved); but they embody certain abstract structures that could, in fact, emerge from other dynamical systems fairly different from the human mind-brain.

To some extent these emergent structures involve “illusions”: “free will” is not as free as we feel it is, our “selves” do not represent ourselves as accurately as we think they do, etc. However, these illusions have arisen as tools for organismic goal-achievement, and it seems that, within the confines of the human brain architecture, we can rid ourselves of them only at the cost of a considerable loss of cognitive efficiency (e.g. meditative states in which illusions are transcended, but not much else gets done).

“Advanced machine consciousness,” to me, indicates the creation of highly intelligent machines (with intelligence equal to or greater than that of humans – and I have no doubt that humans are far from the maximal level of intelligence achievable in this universe, if indeed such a maximum exists), and the study of the emergent structures and dynamics correlated with their most intense subjective experiences. These structures and dynamics will surely have something in common with the structures and dynamics of human consciousness, but one can’t assume them to be identical. Most likely there are certain universal mathematical and conceptual structures associated with intense subjective experience in all intelligent

systems, as well as certain particular structures associated with intense subjective experience in particular classes of intelligent systems (and perhaps differentiating humanlike systems from systems more naturally implemented on networks of von Neumann machines).

2. What is my plan for creating advanced machine consciousness?

If the above view of consciousness is correct, then the task of creating advanced machine consciousness is basically the task of creating mechanical (e.g. computer) systems with high levels of general intelligence. These systems will then (like everything else) have subjective experiences, and certain structures and dynamics will habitually correlate with their more intense subjective experiences.

How then to create artificial general intelligence? I have written a great deal on this topic before and will here give only some vague indications.

First, I conceive general intelligence as the capability to achieve complex goals in complex environments, using limited computational resources.

From this perspective, the basic purpose of a system's intelligence is to, at each time step in the system's life: Enact a procedure so that, according to its best guess, the probabilistic logical implication

Context & Procedure ==> Goals

is true with a high truth value. Here "Context" refers to the current situation as perceived by the system; and a "Procedure" refers to an "internal program" running within the system, that executes a series of behaviors that the system's body knows how to execute. This is an overall perspective on intelligent system function that applies to all intelligent systems whether or not they explicitly use any internal representation of concepts like "probability", "context", "procedure", etc.

The question then becomes how these probabilistic implications are learned and represented. This of course depends on the particular cognitive system in question. What I will talk about here is a system I am developing called the Novamente Cognition Engine (NCE), which also has an open-source variant called OpenCog Prime (OCP).

One way to conceptualize the NCE is to decompose it into five aspects (which of course are not entirely distinct, but still are usefully distinguished):

- Cognitive architecture (the overall design of an AGI system: what parts does it have, how do they connect to each other)
- Knowledge representation (how does the system internally store declarative, procedural and episodic knowledge; and how does it create its own representation for knowledge of these sorts in new domains it encounters)
- Knowledge creation (how does it learn new knowledge of the types mentioned above; and how does it learn how to learn, and so on)
- Instructional methodology (how is it coupled with other systems so as to enable it to gain new knowledge about itself, the world and others)
- Emergent structures and dynamics (which arise from the combination of the four previous)

I now briefly review how these five aspects are handled in the NCE.

The NCE's high-level cognitive architecture is motivated by human cognitive science and is roughly analogous to Stan Franklin's LIDA architecture. It consists of a division into a number of interconnected functional units corresponding to different specialized capabilities such as perception, motor control and language, and also an "attentional focus" unit corresponding to intensive integrative processing.

Within each functional unit, declarative knowledge representation is enabled via an AtomTable software object that contains nodes and links (collectively called Atoms) of various types representing declarative, procedural and episodic knowledge both symbolically and subsymbolically. Each Atom is labeled with a multi-component probabilistic truth value object; and also with a multi-component attention value object indicating its short and long term importance.

Procedural knowledge is represented via program trees in a simple LISP-like language called Combo; and methods exist for translating between Combo and declarative Atoms. Episodic knowledge is represented by the use of Atoms and Combo programs to trigger internal simulations in a UI-free internal virtual world

called Third Life, which may be thought of as the system's "mind's eye" running internal (memory-based or hypothetical) movies.

Each unit also contains a collection of MindAgent objects implementing cognitive, perception or action processes that act on this AtomTable, and/or interact with the outside world.

In addition to a number of specialized learning algorithms associated with particular functional units, the NCE is endowed with two powerful learning mechanisms embedded in MindAgents: the MOSES probabilistic-program-evolution module, and the Probabilistic Logic Networks module for probabilistic logical inference. These are used both to learn procedural and declarative knowledge, and to regulate the attention of the MindAgents as they shift from one focus to another, using an economic attention-allocation mechanism that leads to subtle nonlinear dynamics and associated emergent complexity including spontaneous creative emergence of new concepts, plans, procedures, etc.

Regarding teaching methodology, the NCE has been developed in the context of a physically or virtually embodied approach, which integrates linguistic with nonlinguistic instruction, and also autonomous learning via spontaneous exploration of the physical or virtual world. It is the exploration of the world, and the interaction with other (human) minds in the context of the world, that will, we suggest, allow the system's knowledge-based to adapt in such a way as to give rise to the high-level emergent structures characterizing a human-like mind, and comprising "advanced human consciousness": the phenomenal self, the illusion of will, the theater of reflective awareness, etc.

The NCE is a highly general architecture, and various of its aspects have been implemented and practically utilized to various degrees. One practical application that has recently been constructed utilizing the NCE architecture is the Novamente Pet Brain. This application is the closest we have come, in practice, to applying a Cognition Engine to controlling a human-like autonomous intelligent agent: applying the NCE to controlling a dog-oid virtual-robot living in an online virtual world (the Multiverse or OpenSim worlds, at the moment, though the architecture is extensible to any virtual world). We have also used the NCE in other applications producing more advanced human-like functionalities, e.g. natural language parsing, commonsense deduction based on information extracted from natural language, and biomedical discovery from quantitative and relational databases. But the virtual pet application is a better representation of what we feel is the right path toward powerful artificial general intelligence. I feel there is great potential in the pathway from virtual dogs to virtual talking parrots to virtual babies and eventually to fully-functional virtual adults such as virtual scientists. There is also a strong role for physical robotics along this path, especially as robot simulator and virtual world technology gradually converge, as seems inevitable; but we envision virtual agent control dominating physical robotics as a medium for AGI development in the next decade, due to the far lower cost of development.

The following figure shows two virtual animals controlled by the Pet Brain.



Figure 1. Left: AI-controlled virtual dog in the Second Life, virtual world, learning the behavior "sit." Right: Display of the current emotional and physiological status of an AI-controlled virtual dog in the Multiverse virtual world.

So far the machine consciousness of these virtual pets is pretty primitive! But I feel confident that further developing them according to the NCE/OCP architecture has the potential to yield more and more advanced machine consciousness, up to the human level and beyond.