How to Make a Computer Conscious

Alexei V. Samsonovich (asamsono@gmu.edu)
Krasnow Institute for Advanced Study, George Mason University
4400 University Dr. MS 2A1, Fairfax, VA 22030-444 USA

Introduction

Why do computers lack common sense initiative? What is so special about human brain and consciousness that computers still cannot reproduce? Much brain information processing deals with abstract representations of agents, such as instances of the self and others. Therefore, implementing the right sense of agency in a computer could provide a solution. Here I outline a general approach to the problem of a computer-based implementation of the mind.

Basic Formalism

This approach is based on the formalism of schemas. The notion of a schema used here is more general than any notion of a schema used, e.g., in psychology, or in computer science, or in social sciences. Examples of schemas range from very abstract notions (a thing, an entity, an event, a property, a relation, an agent, an act of learning a schema), to very concrete things (an apple, minus one, red, grasping a pen with the right hand, etc.). In this framework schemas can represent any cognizable thing, including qualia, thoughts, intentions, feelings, and so on. A predicate, a variable, a quantifier, an instruction or even an analog signal can be represented in terms of schemas as well. Elements and components of schemas, as well as complexes made of schemas, can be viewed as schemas on their own. Generally, a schema can be characterized as an abstract model, a template or a prototype that can be bound to a particular content. Definitions of particular schemas must specify semantics, syntax, pragmatics and dynamics. Schemas can be represented in a computer symbolically in a standardized format with the structure of a nested list.

When an instance of a schema is bound to a particular content, it forms a state. Schemas and states are dynamical objects in this framework. They evolve in time according to rules defined by schemas. E.g., states can be initiated, executed and terminated; dynamics of a state may affect other states and schemas. A chart is a special state that represents a mental perspective of an agent (e.g., I-Now, I-Yesterday, He-Imagined). When a state is bound to a chart, it represents a mental state. Perspectives change with time according to the flow of the subjective time: I-Now becomes I-Previous, and so on. In addition to perspectives, charts have internal dimensions that determine attitudes of the associated states: e.g., I-Now may contain a belief about a past or a future. The general idea of this framework based on schemas is not new; however, its known analogs (e.g., OPS, SOAR, ACT-R) lack generality and do not implement a proper concept of the self possessing free will.

Implementing Free Will and The Self

Charts labeled "I-…" represent instances of the self of the virtual individual, including cognitive and metacognitive perspectives. Other charts may represent mental simulations of "third persons": in other words, a "theory of mind". Only the content of I-Now represents the current content of consciousness in this framework. Mental states in I-Now have special privileges: e.g., they can directly determine scheduling of voluntary acts and may access all other charts.

A distinguishing feature of this framework is that it is based on the fundamental properties of the human self (*error fundamentalis*: Nadel & Samsonovich, 2002), including its uniqueness, its localization in a particular context, its indivisibility, its self-consistency over time, and finally, its apparent free will and self-awareness that are present in the contents of consciousness at any instance.

These properties are introduced via the design of the system or via dynamical constraints. They guarantee certain degree of coherence in system's behavior. Their products are the emergence and the maintenance of a unique working scenario (i.e., a consistent sequence of charts leading from I-Now to I-Goal) and the generation of voluntary actions. An action is considered voluntary, iff it results from an intention of "I" (represented by a mental state in I-Now), is consistent with the working scenario and is not biased by any additional factor. Intentions are selected among the available ideas based on their fitness into the working scenario, and ideas (i.e., mental states in I-Now that represent feasible actions) result from the activity of states.

Previously active charts may disappear with time together with their contents – or be remembered, thus contributing to the episodic memory and/or to the system of values, i.e., the memory of dreams and goals. In contrast, the semantic memory is represented by the set of schemas. Creation of new schemas is controlled by the learning schemas and may be supervised or not. A complete system should be able to work interactively with a human instructor, using a specially designed language. The range of interaction paradigms may include listening to and completing stories, learning how to play or playing games, learning languages or scientific disciplines, operating a robot in a virtual environment, simulating virtual characters, designing or training other systems of the same kind, etc. Primary applications of this framework should relate to natural language understanding.

References

Nadel, L., & Samsonovich, A. (2002). The conscious self. To appear in: S. Jess (Ed.). *Brain, Mind & Consciousness*. CA: University Press.