Todd Moody's Zombies

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1997 Feb 28, 6:24 a.m.

Abstract

From the AI point of view, consciousness must be regarded as a collection of interacting processes rather than the unitary object of much philosophical speculation. We ask what kinds of propositions and other entities need to be designed for consciousness to be useful to an animal or a machine. We thereby assert that human consciousness is useful to human functioning and not just and epiphenomenon. Zombies in the sense of Todd Moody's article are merely the victims of Moody's prejudices. To behave like humans, zombies will need what Moody might call *pseudo-consciousness*, but useful pseudo-consciousness will share all the observable qualities of human consciousness including what the zombie will be able to report. Robots will require a *pseudo-consciousness* with many of the intellectual qualities of human consciousness but will function successfully with few if any human emotional conscious qualities if that is how we choose to build them.

1 Introduction

From the AI point of view, consciousness must be regarded as a collection of interacting processes rather than the unitary object of much philosophical speculation. We ask what kinds of propositions and other entities need to be designed for consciousness to be useful to an animal or a machine. We thereby assert that human consciousness is useful to human functioning and not just and epiphenomenon. Zombies in the sense of Todd Moody's article are merely the victims of Moody's prejudices. To behave like humans, zombies will need what Moody might call *pseudo-consciousness*, but useful pseudo-consciousness will share all the observable qualities of human consciousness including what the zombie will be able to report. Robots will require a *pseudo-consciousness* with many of the intellectual qualities of human consciousness but will function successfully with few if any human emotional conscious qualities if that is how we choose to build them.

Such is an AI doctrine on the subject. We now must ask what are the specific processes that make up the consciousness necessary for successful robots and the additional processes required should we want them to imitate humans. Many aspects of intelligent behavior do not require anything like a human level of consciousness, and hardly any AI systems built so far have any. For this reason the following remarks are somewhat speculative and are more stimulated by people like the Dreyfuses and Penrose who deny the possibility of robot consciousness than by any features of existing programs.

We regard consciousness as a subset of the memory of an animal or machine distinguished by the fact that many processes involve only those elements of memory that are in consciousness. The elements of memory include propositions (like sentences) and other entities We may divide our consideration into *basic consciousness* and *consciousness of self*.

2 Basic Consciousness

Here are some of the elements of basic consciousness.

propositions The propositions of basic consciousness are about the world and not about the system's thoughts. There is a gray area, and, for example, a proposition that the system is hungry can be looked at the other way.

images of scenes and objects These may be either remembered images or images of objects currently being sensed. By image, I do not mean merely two dimensional visual images such as those projected on the retina. Included are auditory images and three dimensional images of objects. The auditory images of speech are transformed by filters characteristic of the hearer's understanding of the language. The images of three dimensional objects involve vision, touch and also experience with the particular kind of object.

Much more can be said about images, but it is inessential for this review, except to make the point that the actual details are important in understanding consciousness.

3 Consciousness of Self

[McC95] discusses the kinds of consciousness of its own mental processes a robot will require in order to behave intelligently. Here are a few of them.

1. Keeping a journal of physical and intellectual events so it can refer to its past beliefs, observations and actions.

2. Observing its goal structure and forming sentences about it. Notice that merely having a stack of subgoals doesn't achieve this unless the stack is observable and not merely obeyable.

3. The robot may *intend* to perform a certain action. It may later infer that certain possibilities are irrelevant in view of its intentions. This requires the ability to observe intentions.

4. Observing how it arrived at its current beliefs. Most of the important beliefs of the system will have been obtained by nonmonotonic reasoning, and therefore are usually uncertain. It will need to maintain a critical view of these beliefs, i.e. believe meta-sentences about them that will aid in revising them when new information warrants doing so. It will presumably be useful to maintain a pedigree for each belief of the system so that it can be revised if its logical ancestors are revised. *Reason maintenance systems* maintain the pedigrees but not in the form of sentences that can be used in reasoning. Neither do they have introspective subroutines that can observe the pedigrees and generate sentences about them.

5. Not only pedigrees of beliefs but other auxiliary information should either be represented as sentences or be observable in such a way as

to give rise to sentences. Thus a system should be able to answer the questions: "Why do I believe p?" or alternatively "Why don't I believe p?".

6. Regarding its entire mental state up to the present as an object, i.e. a context. [McC93] discusses contexts as formal objects. The ability to *transcend* one's present context and think about it as an object is an important form of introspection, especially when we compare human and machine intelligence as Roger Penrose (1994) and other philosophical AI critics do.

7. Knowing what goals it can currently achieve and what its choices are for action. We claim that the ability to understand one's own choices constitutes *free will*. The subject is discussed in detail in [MH69].

Taken together these requirements for successful human-level goal achieving behavior amount to a substantial fraction of human consciousness. A human emotional structure is not required for robots.

4 Moody Zombies

Moody isn't consistent in his description of zombies. On the page 1 they behave like humans. On page 3 they express puzzlement about human consciousness. Wouldn't a real Moody zombie behave as though it understood as much about consciousness as Moody does?

References

[McC93] John McCarthy. Notes on formalizing context. In *IJCAI-93*, 1993. Available on http://www-formal.stanford.edu/jmc/.

[McC95] John McCarthy. Making robots conscious of their mental states. 1995. to appear, available on http://www-formal.stanford.edu/jmc/.

[MH69] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

/@sail.stanford.edu:/u/jmc/e95/zombie1.tex: begun 1995 Jul 30, latexed 1997 Feb 28 at 6:24 a.m.