Artificial Intelligence in Medicine (2008) 44, 139—146

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Artificial consciousness: Hazardous questions (and answers)

Giorgio Buttazzo *

Scuola Superiore Sant'Anna, Via Moruzzi 1, 56124 Pisa, Italy

Received 19 February 2008; received in revised form 30 June 2008; accepted 3 July 2008

KEYWORDS
Artificial consciousness;
Conscious machines;
Artificial neural networks;
Artificial beings;
Self-aware systems

Summary If human consciousness is the result of complex neural electro-chemical interactions occurring in the brain, the question of whether a machine can ever become self-aware could be a matter of time: the time necessary to fully understand the functional behavior of the brain structure, develop a mathematical model of it, and implement an artificial system capable of working according to such a model.

This paper addresses several issues related to the possibility of developing a conscious artificial brain. A number of hazardous questions are posed to the reader, each addressing a specific technical or philosophical issue, which is discussed and developed in a form of a hazardous answer.

1. Introduction

The idea of building a smart machine capable of competing with human intelligence has always been a dream of some computer scientists since the early establishment of computer technology [1,2]. Today, given the successful results of artificial intelligence and neural computing, and considering the current pace of computer evolution, such a possibility seems to be more concrete than ever, and someone believes that, in the near future, machines will exceed human intelligence and eventually will develop a mind. Talking about artificial consciousness, however, gives rise to many philosophical issues [3]. Are computers thinking, or are they just calculating? Conversely, are human beings thinking, or are they just calculating? Is consciousness a prerogative of human beings? Does it depend on the material the brain is made of or can it be replicated using a different hardware? Answering these questions is not easy, since it requires moving along the edges of several different disciplines, such as computer science, neurophysiology, philosophy, and religion. Nevertheless, many people believe that artificial consciousness is possible and that in the future it will emerge in complex computing machines. In the rest of this paper, a number of provocative questions are posed to the reader, each addressing a specific technical or philosophical issue, which is discussed and developed in a form of a hazardous answer.

2. What is artificial consciousness?

As Tagliasco pointed out [4], the term "artificial" is often used in two different meanings. In a first form,

* Tel.: +39 050 882012; fax: +39 050 882003.
E-mail address: buttazzo@sssup.it.

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2008.07.004

the term "artificial" denotes a human artifact that replicates or simulates a real thing. For example, an artificial flower is something that appears as a flower, may have the same shape and colors, but it is very different in terms of materials and functions. In this sense, an artificial thing is a very simplistic version of its real counterpart. In some other cases, the term "artificial" is used to denote a genuine physical phenomenon reproduced using a human made device. For example, "artificial light" denotes an electromagnetic wave produced by a human made device, like a bulb or a led. Depending on the meaning we associate with the term "artificial", we can distinguish two types of artificial consciousness, as proposed by Holland [5,6]:

- *Weak artificial consciousness*: It is a simulation of a conscious behavior. It can be implemented as a smart program that simulates the behavior of a conscious being at some level of detail, without understanding the mechanisms that generate consciousness.
- *Strong artificial consciousness*: It refers to a real conscious thinking emerging from a complex computing machine (artificial brain). In this case, the main difference with respect to the natural counterpart depends on the hardware that generates the process.

In this paper we are interested in Strong Artificial Consciousness and we will speculate on the possibility that such a form of consciousness can emerge in a complex computing system.

3. How can we verify consciousness?

In 1950, the computer science pioneer Alan Turing posed a similar problem but concerning intelligence. In order to establish whether a machine can or cannot be considered intelligent as a human, he proposed a famous test, known as the Turing test: there are two keyboards, one connected to a computer, the other leads to a person. An examiner types in questions on any topic he likes; both the computer and the human type back responses that the examiner reads on the respective computer screen. If the examiner cannot reliably determine which was the person and which the machine, then we say the machine has passed the Turing test.

In 1990, the Turing Test received its first formal acknowledgement from Hugh Loebner (a New York philanthropist) and the Cambridge Center for Behavioral Studies (Cambridge, MA), which established the Loebner Prize Competition in Artificial Intelligence [7]. Loebner pledged a prize of $100,000 for the first computer whose responses were indistinguishable from those of a human. The first competition was held at the Computer Museum of Boston in November 1991. For some years, the contest was constrained to a single narrow topic, but the most recent competitions, since 1998, did not limit the scope of questioning. Each judge, after the conversation, gives a score from 1 to 10 to evaluate the interlocutor, where 1 means human and 10 computer. So far, no computer has given responses totally indistinguishable from a human, but every year scores are getting closer to five in the average [8]. Today, the Turing test can be passed by a computer only if we restrict the interaction on very specific topics, as chess.

On 11 May 1997 (3:00 p.m. eastern time), for the first time in the history, a computer named Deep Blue beat world chess champion Garry Kasparov, 3.5—2.5. As all actual computers, however, Deep Blue does not understand chess, since it just applies some rules to find a move that leads to a better position, according to an evaluation criterion programmed by chess experts.

Claude Shannon estimated that in a chess game the search space includes about 10^{120} possible positions. Deep Blue was able to analyze 200 million (2×10^8) positions per second. Exploring the entire search space for Deep Blue would therefore take about 5×10^{111} s, which is about 10^{95} billions of years. Nevertheless, Deep Blue victory can be attributable to its speed combined with a smart search algorithm, able to account for positional and material advantage. In other words, computer superiority was due to brute force, rather than sophisticated machine intelligence.

In spite of that, in many interviews during and after the match, Kasparov expressed doubts he was playing against the computer and sometime he felt like playing against a human. In some situation, he also appreciated the beauty of the moves done by the machine, as if it was driven by intention, rather than by brute force. Thus, if we accept Turing's view, we can say that Deep Blue plays chess in an intelligent way, but we can also claim that it does not understand the meaning of his moves, as a television set does not understand the meaning of the images it displays.

Besides chess, there are other domains in which computers are reaching human ability, and their number is increasing every year. In music, for example, there are many commercial programs that can create melodic lines or even entire songs according to specific styles, ranging from Bach to jazz. There are also programs that generate great solos on top of a given chord sequence, emulating jazz masters, like Charlie Parker and Miles Davis, much better