Artificial Intelligence: The Mind-Body Problem Revisited in the Computer Age

An informal introduction

Haldun M. Özaktaş
Bilkent University, Ankara, Turkey

Lecture notes for GE 301: Science, Technology, and Society, Autumn 1997 version

Abstract

Humans design and build artificial systems which to some extent exhibit autonomous behavior. Extrapolating technological advances, will it be ultimately possible to build artificial systems which can be said to possess intelligence, feelings, consciousness? What do these terms mean anyway? The issue is whether human qualities such as "intelligence" and "feelings" can or cannot be reduced to mechanistic or algorithmic terms. While the advocates of both sides of the issue cannot come up with definitive arguments, it is interesting to observe that this conflict reflects long standing divides in Western philosophy.

1 Introduction

The subject matter of these notes is usually found under the title "Philosophical Implications of Artificial Intelligence."

The first question we must ask is "What can a computer do?" Or perhaps, we should ask, as in Hubert Dreyfus' influential book, "What computers can't do?" We know that computers can execute algorithms, that is, well defined procedures. Computers can clearly carry out anything we can write an algorithm for.

Now, let us ask ourselves whether computers can engage in creative activity of the sort we usually associate with human intelligence. Can they develop a new branch of mathematics, improve our understanding of the physical world under their own initiative, or write poetry? Based on what we just said above, these questions boil down to whether these activities can be reduced to a well defined procedure. When we engage in creative or intelligent behavior, are we merely executing an algorithm stored in our brains?

If all human actions can be reduced to well defined procedures, then it means that in principle a computer can be made to indistinguishably simulate a human being, at least mentally and verbally. Thus, a computer would be able to simulate a human being in love. A somewhat more subtle issue is whether in this case we would say that the computer *is* in love, whether it actually feels the same way as somebody in love does. In more general terms, are what we call feelings and consciousness merely an emergent manifestation, an illusion arising as the result of our brains carrying out a very complicated algorithm?

Let us first make precise the notion of a computer indistinguishably simulating a human being. An *operational* test of indistinguishability was first put forward by Alan Turing and is known as the Turing test (the teletype scenario). In the Turing test, we view humans and computers alike as a black box. If they turn out to be identical in terms of input/ouput relations, (that is, we cannot say which is the human and which the computer), the computer is said to "pass" the test.

One must distinguish two issues. First, can a computer be built that at least as far as verbal inputs/outputs are concerned, can actually pass the Turing test even after exhaustive probing? If not, that would mean that there are some things that humans can do which cannot be reduced to a computer program.

But now, let us for the moment assume that they can pass the test. We are now faced with a difficult question. If the computer exhibits identical input/output behavior to a human being, must we state that the computer is intelligent, or that it is in love.

Before further elaborating on this point, it is worth noting that this way of thinking is associated with the school of human sciences known as *behaviorism.* According to behaviorists, for sociology and psychology to be a proper science, one must do away with wishy washy concepts like internal states of the mind, consciousness, the subconscious, the spirit, and so forth and concentrate on determining, most probably by experiment, the input/output relations of human beings, social systems, rats, whatever. Since only input/output relations are what are visible, only they are the materials of a proper scientific method. A typical behaviorist experiment would be putting people in a room, facing them with a certain situation, and measuring how they behave. Although behaviorism has dominated US and thus world sociology for long, it is now acknowledged that it is of value for only certains kinds of problems and is considered inadequate from a general standpoint.

2 What does it mean to feel pain?

Let us consider a concrete example and ask "What does it mean to feel pain or pleasure?" The importance of feedback in living systems was recognized early and is generally referred to as *homeostasis.* Living things adapt to their environment in order to maintain internal stability, such as control temperature. The similarity of this operation with that of the archetypical feedback system, the thermostat, led some people in the first half of the century to attach central importance to the role of feedback in living systems. It was considered the essence of complex behavior and life. This led to the General Systems Theory movement of von Bertallanfy in which the paradigm of all things was essentially a coupled set of possibly nonlinear differential equations, which can be interpreted as a complicated feedback network. This movement was also closely linked to Cybernetics. Overall, the idea was to view living beings as automatic control systems.

So after this aside, let us get back to our concrete example. Say I touch something hot and pull back my hand. That is feedback. An automatic system which does the same can be conveniently built. We say we "feel" pain. Is this "to feel" really something tangible, or is it simply a fictitious label which really means "My feedback system was just activated." We are hurt and we do not like it, but perhaps the act of not liking something is no less fictitious, no less an illusion than that of feeling pain. Most people would not admit that the machine feels pain, but since it is behaviorally identical to us, perhaps our situation is not different after all. Or, on the other hand, if we insist on defining our situation as "being in pain," must we then not have to accept that the machine feels as well, and that feeling is nothing more than what happens to the machine.

We see that the behaviorist stance forces us to either accept that both we and the machine feel pain, or that we both don't. It is simply an ad hoc definition whether we call this situation pain. Perhaps then, there is no point in discussing whether we are in pain or not, perhaps this notion of "feelings" is unscientific and should be abandoned. This line of thought is associated with the *logical positivism* of the Vienna School, first half of the century. According to them their were two kinds of truths. Logical truths, which were tautologically true, and empirical truths that were experimentally verified. Anything that could not be set up for an operational verification (or later under Popper, falsification), was unscientific and meaningless to discuss.

Thought experiments similar to that above about pain lead us to questions such as "If the machine can simulate intelligence, is it intelligent?" or "If the machine can simulate consciousness, is it conscious?"

Let us be sure we do not confuse the two questions: Can the machine behave as, for example, it understands something it reads? Even if it does, does it mean it actually understands?

Searle has come forward with the thought experiment of the "Chinese room" in an attempt to prove the negative. Say we have a well defined procedure of understanding Chinese stories and that I sitting in a room carry our this procedure so that I am able to answer any questions about the story correctly. Does that mean I understand the story? No, Searle says. But Searle is missing the point. The human operator is not supposed to understand, rather the system "understands." In other words, "to understand" is nothing but what the system does. Or perhaps, the concept of "to understand" is meaningless.

3 What is a mind anyway?

After having discussed some of the central issues somewhat randomly, let us now try to isolate a number of viewpoints regarding the status of the human mind.

3.1 Vitalism

There is some "essence" which makes a human a human. Thus a computer, no matter what it does, cannot have a mind or feelings. "It is just a machine."

This viewpoint may or may not be of religious origin, but it clearly assumes that there are some things that are significantly beyond current "scientific" understanding.

3.2 Mechanistic-reductionism

Rooted in Western philosophical tradition strengthened by the rise of empiricism and culminating in Newtonian mechanics. The universe is a deterministic/mechanistic system of interacting atoms, including all living beings. (It is interesting to note that this is still the unconscious model of reality shared by most scientists and engineers. This also leads directly to the issue of free will etc., but I will avoid that.)

Originally, at the time of Descartes, the above model of the universe was supplemented by the human mind, considered to be of a wholly different nature. (Animals were assumed to be mere automatons.) This is referred to as Cartesian dualism. It was a serious problem how this intangible mind could possibly interact with the mechanistic body of a human. Various solutions were offered. One called "parallelism," claimed that the mind and body did not interact at all, and that their harmony was a result of divine design. (The example of two noninteracting clocks is often given to make this point.)

Although Descartes could not shake off the notion of the mind stemming from monotheist origins, current AI flag wavers like Marvin Minksy have no problems stating that we are no different than our physical environment, and are merely computers made of flesh, which is also made of atoms, and that our status is no different than that of artificially built computers. The notion of the "mind" is either dropped, or it is claimed that computers can have as much of it as we do.

3.3 Non-algorithmic paradigms

This is the view of the hard-headed scientist who cannot accept the vitalist viewpoint, but is not willing to believe that human activity can be reduced to well defined procedures. Such a person believes that we humans, just as everything else, are organisms or systems made of the same stuff as the rest of the universe and are governed by the same physical laws. Thus, if we are made of a finite number of atoms, it is conceivable that a similar number of atoms can be put together deliberately to create an "artificial" thing, which will be just like a human in every way.

(Of course, this is just what people do when they make babies. Thinking further, one is led to question what it means for something to be artificial. For instance, is a beehive artificial?)

However, such a person would say, perhaps there are physical laws which we do not sufficiently understand now. Perhaps, although governed by the laws of this universe, the human mind is non-algorithmic, that its operations and input/output relations cannot be reduced to a well defined procedure. Note that this does not exclude the possibility of an artificial system that can be just like a human being. However, such a machine will not be equivalent to a Turing machine, it will not be algorithmic in the sense that all computers today are.

There are some arguments that suggest that the human mind might be non-algorithmic, or that there are at least some non-algorithmicity in the universe. Gödel's theorem tells us that there are some problems for which we cannot write an algorithm for, that would tell us whether they are true or false. But there are examples in which although this is the case, the human observer can actually determine whether it is true or false. (We just don't have a well defined procedure that guarantees that we would be able to do so.) Thus, our brains must have some non-algorithmic elements.

That is all nice, but it is not so easy to imagine something which is not algorithmic. Our common sense is so shaped by the determinism of classical physics, that we are unable to think of a process which cannot be modeled on the basis of an algorithm. Indeed, it seems that everything in the universe can be simulated by a cellular automata, differential equations, ultimately a well defined procedure.

4 What could be the new laws of physics supporting non-algorithmic behavior?

What could be these new physical laws which we do not know and which defy such reduction?

Weak candidates are the neural network-connectionist paradigms, with their "emergent" behavior etc. However, there is nothing to make us think that a neural network is fundamentally different than existing computational systems.

A major contender is chaos and complexity theory, which has revealed the fundamentally different behavior of even relatively complex systems. It is yet to be seen what they have to say about this issue.

Finally, I would like to mention the possibility of new interpretations of quantum theory, especially in relation to gravitational effects. This possibility was raised by Roger Penrose. He suggests that perhaps beyond a critical level of complexity a qualitatively different type of non-algorithmic behavior sets in. He speculates that this critical level is exceed in our brains, but not in current computers.

5 Conclusion

There are some issues and problems that have been around ever since. The way these problems are posed, however, reflects the time and place of those pondering them. Thus it is only natural that in an age where the computer is of such prominence, questions about the mind and body, free will, determinism, the status of human beings in the universe, etc. come up in such a context.