Artificial Intelligence and Consciousness

Goutam Paul

Department of Computer Science
State University of New York at Albany
1400 Washington Avenue, Albany, NY 12222

goutam@cs.albany.edu

Abstract

The mystery of consciousness has enthralled human beings from time immemorial. The formal study of consciousness had been restricted mainly to philosophers and logicians in the past. However, with the development of fields like Computers and Cognitive Science, people from different disciplines are studying this subject and shedding new light upon it. Research in Artificial Intelligence has given rise to interesting questions such as: "Can human beings create machines who can think and act like humans?" This paper tries to present an informal overview of Artificial Intelligence, analyze the relationship between intelligence and consciousness, and finally discuss the possibility of artificial intelligence and consciousness.

1. Introduction

Artificial Intelligence (AI) is one of the newest disciplines which attempts to understand intelligent entities. One reason we study AI is to understand ourselves better. Fields like Philosophy and Psychology also try to do the same, but the difference is that AI not only tries to understand human intelligence, it tries to build human-like intelligent entities as well. Computers with human-level intelligence would definitely have a tremendous influence on our daily lives, maybe not exactly as depicted by many science fiction movies, but in similar and many other ways. AI is a multidisciplinary field and apart from Computer Science, it takes concepts from Mathematics, Logic, Probability Theory, Statistics, Control Systems, Information Theory, Philosophy, Psychology, Neurobiology and many other disciplines. However, in this paper, we will have a non-technical view of the subject and try to analyze from philosophical and intuitive notions the possibility of machines with human-like intelligence and consciousness.

2. Four Approaches to AI

There are different approaches to the definition of AI. One dichotomy is *thought* versus *behavior*. Another is *human-like intelligence* versus *ideal intelligence* (also called *rationality*)[§]. The table below from a modern AI text book [1] summarizes these two dichotomies as follows:

	Human	**Ideal**
Thought	Systems that think like humans	Systems that think rationally
Behavior	Systems that act like humans	Systems that act rationally

Table 1: Different Definitions of Artificial Intelligence

[§] It does not assume that human beings are irrational. Rather human beings do not "always" act rationally.

"Acting humanly" takes the Turing Test approach [2]. As part of his argument, Alan Turing (1912-1954) put forward the idea of an "imitation game", in which a human interrogator interacts with a human being and a computer under conditions where the interrogator would not know which is which. For example, we can keep a human being and a computer in two separate closed door rooms, and the interrogator can communicate with them entirely by exchange of textual messages. Turing argued that if the interrogator could not distinguish them by questioning, then it would be reasonable to call the computer intelligent.

"Thinking humanly" takes the cognitive modeling approach which gets inside the actual workings of the human mind. Cognitive Science is an interdisciplinary field. It brings together computer models from AI and experimental techniques from Psychology in order to construct testable theories of the workings of the human mind.

"Thinking rationally" takes the laws of thought (also called *Logic*) approach, first initiated by the Greek philosopher Aristotle (384-322 B.C.). A famous example of logical reasoning is: "Socrates is a man; all men are mortal; therefore Socrates is mortal."

"Acting rationally" takes the rational agent approach. An *agent* is something that has a set of sensors to observe the *state* of its environment, and a set of pre-defined *actions* it can perform to change one state into another. For example, a mobile robot may have sensors such as camera and sonars and actions such as "move forward" and "turn." A more sophisticated example is agent Smith in the popular movie "Matrix." Currently, most AI research is concerned with the study and construction of rational agents.

3. Looking Back in History

Though AI is a recent field of study, the first step towards AI by any human being took place when Socrates (469-399 B.C.) was seeking an *algorithm* (i.e. a procedure or recipe, analogous to a computer program) to distinguish between goodness and badness. Socrates' disciple Plato (427-347 B.C.) and grand-disciple Aristotle (384-322 B.C.) formulated the laws governing the rational part of the mind (we may call this rational part as intelligence). Aristotle laid the foundation of Logic which allows one to generate conclusions in a step-by-step way starting from a fixed set of axioms or assumptions. This gave the human society hope to understand mind by mathematics. Later, George Boole (1815-1864) mathematized Aristotle's system of reasoning and gave rise to Symbolic Logic as we understand it today.

However, French philosopher and mathematician Rene' Descartes (1596-1650) pointed out a problem with purely mechanical and physical conception of mind: it leaves no room for free-will. He said that there is a part of the mind that is outside of nature, exempt from the physical laws.

Kenneth Craik (1914-1945) ushered the area of Cognitive Psychology and claimed that belief and reasoning are as scientific components of human behavior, as are temperature and pressure of gases.

Alan Turing (1912-1954) formalized the notion of modern-day computers and gave rigorous mathematical characterizations of how a computer program works. Von Neumann (1903-1957) pioneered Decision Theory which gave a theory to distinguish between good and bad actions as dreamt of by Socrates.

In 1943, neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper on how neurons in the brain might work. They modeled a simple neural network using electrical circuits which began the sub-field of neural computing in the AI discipline.

4. The Present and the Future

The second half of the twentieth century saw revolutions one after another in the field of AI. To trace all of them would require another complete paper. In short, AI became state of the art from game playing to automatic traffic control system, from forensic analysis to medical diagnosis, from weather prediction to trading agent in stock markets. People have developed programs that successfully learn to recognize spoken words, predict recovery rates of patients, detect fraudulent use of credit cards, drive autonomous vehicles on public highways, play games at levels approaching the performance of human world champions, etc. NASA has been using an AI system to classify celestial objects in their Sky Survey. So far, artificial neural networks have not even come close to modeling the complexity of the brain, but they have shown to be good at problems which are easy for a human, but difficult for a traditional computer, such as recognizing images and making predictions based on past knowledge.

The future is guided by the past and the present. The way current research activities are going, the coming years will definitely see AI flooding every sphere of our lives. We may have robots as domestic servants, AI systems doing operations on patients and many more things, or we may even have a soccer team with mixed humans and artificial intelligent agents. Science fiction movies show many marvelous achievements of AI. Some of them would obviously come into being someday; it is just a matter of time.

5. Limitations of AI: Machine Mind versus Human Mind

We should realize that whatever is imagined and dreamt of in AI, all of that cannot happen. There are some inevitable and inherent limitations. One such limitation that shocked the world was discovered by the famous logician Kurt Gödel (1906-1978) in 1931 in his two *Incompleteness Theorems* [3]. Mathematically, they are highly rigorous, but philosophically they are very simple and even a layman can understand and appreciate them. We will take this layman's view here.

Any scientific theory begins with a set of assumptions called *axioms*, which are taken as self-evident truths. All theorems and results that are proven subsequently rely upon these axioms. For example, the knowledge of geometry starts with the assumption of existence of a *point*. The knowledge of Physics starts with the assumption that there are three things: *matter*, *space* and *time*. The knowledge of Mathematics and Computers is based on the assumption of the *numbers* such as 1, 2, 3, … etc.

Informally speaking, Gödel's first incompleteness theorem says, "Given any axiom system, there will always be some true statement which the system will not be able to prove." But how does this limit the power of AI? The argument is simple. If we believe that we can understand the human mind in terms of mathematical and logical analysis, then by Gödel's first theorem there will always be some truth about our mind which we will never be able to know! If we cannot completely understand our own mind and intelligence, how can we develop an intelligent being exactly like us?

Gödel's second incompleteness theorem deals with *consistency*. An axiom system is called consistent if the system proves a statement to be either true or false, but not both. As a simple analogy, if we say, "It is raining," then we are consistent, because the fact that "It is raining" is either true or false, but not both. Similarly, if we say, "It is not raining," then also we are consistent. But if we say, "It is both raining and not raining," then we are inconsistent.

Informally, Gödel's second incompleteness theorem says, "If an axiom system is consistent, then it cannot verify its own consistency." What is its implication in AI? On one hand this theorem means that if we design a robot, then that robot cannot verify its own consistency, because after all the robot is a product of axiomatic science. On the other hand, this theorem means that the human mind is not just an axiom system. Had it been so, we would not be able to verify our own consistency. Whereas in reality, we do know that we are consistent. For example, nobody says: "I am having a headache and not having a headache" at the same time. If we believe that we are inconsistent, then we would rather stop thinking altogether, and go crazy. Because we know our own consistency, our mind cannot be purely mechanical, there has to be some "spirit" part to it as conjectured by Rene' Descartes (see section 3 above).

6. Body, Mind, Intelligence, and Consciousness

So far, we have used the words "mind" and "intelligence" interchangeably. However, if we analyze carefully, they stand out to be distinct.

The grossest level of existence of a living being, like the human, is the body which is made up of the sense organs. The fact that the mind is distinct from and subtler than the body is not hard to understand. Often times, we feel pain in our mind though there may not be any pain in the body. Again, there may be a painful wound in the body, but if the mind is engaged in something pleasing and enjoyable, then we may forget the wound temporarily and feel no pain at all. The mind is simply a repository of thoughts and feelings. However, the intelligence is subtler than the mind. Intelligence can be defined as that entity which has the power to discriminate between right and wrong actions (or, between rational and irrational actions, as per the terminology of section 1). As an example, suppose a doctor has given some bitter syrup to a patient. Now, the patient's mind may be totally averse to taking such a medicine, the mind may say, "Do not take the medicine," whereas his or her intelligence says, "Well, if you do not take the medicine, your disease won't be cured." Someone might argue that it is a dichotomy in the mind itself – the rational part of the mind versus the irrational part of the mind. But the reason we prefer to say that the intelligence is an entity separate from the mind, is the controlling power of the intelligence over the mind. There seems to be a hierarchy, as the mind has the power to control the body, and the intelligence has the power to control the mind. The mind functions based on emotions and so if we do something at the spur of the moment, driven by our instincts, often times we commit mistakes. It is the intelligence that has the power of discrimination and rational decision. When we say, "Take a decision by the brain, not by the heart," we actually mean: "Take a decision by the rationality of intelligence, not by the emotions of mind." However this does not prove that heart is the seat of the mind and the brain is the seat of the intelligence; that is altogether a different topic and beyond the scope of our current discussion.

Consciousness is an entity which is the subtlest of all. It is beyond body, mind, and intelligence. For example, suppose a patient is under a coma. He or she does not think or feel anything, and so his or her mind is not active. He or she does not do any logical analysis and make any decision, and so his or her intelligence is not active either. But he or she is still alive, and so the life-force or the consciousness is fully present in him or her. In short, consciousness is the symptom of life. People have tried to link consciousness with the brain, but with no success [4]. Mind and intelligence may have some connection with the brain, as different types of living beings have different levels of intelligence. But consciousness does not seem to be merely a product of the nervous system or the brain. Trees do not have any nervous system or brain, but they have consciousness. Whether a single living cell has a mind or intelligence may raise disputes, but it undoubtedly possesses consciousness.

Consciousness can also be related to what is called one's ego or identity. For example, a person may grow from childhood to old age, undergoing changes in his or her body, mind (in terms of thoughts, feelings), and intelligence (hopefully he or she would be more mature and more rational), but the identity of the person remains unchanged. Each one of us knows that it's "me" and nobody else, and the same "I" that existed ten years back is existing now. This "I" or "ego" can be viewed as consciousness. When some philosophy talks about the soul, they refer to this very consciousness [5].

7. Superiority of Consciousness

The fact that intelligence is higher than the mind and it is associated with rationality does not guarantee that decisions taken by the intelligence are always right. Just like in Logic, even if every step in the argument is perfect, if the initial assumptions or axioms happen to be wrong, then all the conclusions would be wrong. As an example, suppose we have a perfect computer program, but the input to this program is wrong (say, not in the format that the program assumes its input to be). Then even if every step of execution of the program is perfect, the output will be wrong.

Since axioms are taken for granted and cannot be proven by rationality or intelligence, the only way to choose between right and wrong sets of axioms is by applying proper consciousness. In all fields of science, this is how the axioms originate – they emerge from the core of human consciousness, and then by applying intelligence upon those axioms human beings develop the subsequent theory or justify observational results.

We can also understand the different levels of subtlety of body, mind, intelligence, and consciousness by the way we feel satisfaction about each of these entities. For example, sexual activity may satisfy one's body (and perhaps partially satisfy one's mind too, as the mind is just the next level above the body and is thus in direct touch with the body). But it does not satisfy the intelligence, what to speak of satisfying the consciousness. Similar things happen when we are hungry and eat some food. It satisfies the body (and maybe partially the mind), but nothing beyond. On the other hand, when we see some beautiful scenery in nature or listen to melodious music, our mind is satisfied and also our fatigued body may be relaxed and re-energized. The intelligence is satisfied when, for example, we solve a hard mathematical problem, or compose a poem, or win a debate. Since the intelligence is higher than the mind, when the intelligence is satisfied, the mind is also satisfied, but the consciousness which is higher than the intelligence is not satisfied. However, when we do sacrifice out of love, and when we feel for other's sufferings and try to do service to others, our consciousness is satisfied. Since the consciousness is the highest of all, naturally the intelligence and the mind are also satisfied along with it.

As an analogy to the hierarchy of body, mind, intelligence, and consciousness, we can imagine a chariot. The body is like the horse, and the mind is like the rope tied to the horse. The intelligence is the driver who holds the rope in his or her hand, and who has the power to control the rope. The consciousness is like the passenger in the chariot, who directs the driver. The chariot driver would drive the horses in the right path, provided the passenger is able to instruct him or her properly. Similarly, in order for the intelligence to take the right course of action and engage the mind and the senses in proper activities, it has to be properly guided by the consciousness.

8. Machine Intelligence and Consciousness

So far we have talked about AI in terms of designing intelligent machines. But now that we have discussed the relationship between intelligence and consciousness, the next question is: "Can we design conscious machines, i.e. machines which can identify a unique “me” in them?” As far as the current research goes, it does not seem feasible. In that sense, the term Artificial Intelligence is co-incidentally appropriate to the subject matter - people have not coined the term Artificial Consciousness maybe because of the seeming impossibility of its existence. But is there any justification why there cannot be conscious machines? It seems there is.

Intelligence understood as rationality can be defined and analyzed by mathematics and logic to some extent. And the whole of Computer Science and AI is based primarily on mathematics and logic. However, when it comes to consciousness, mathematics ends and philosophy begins. Scientists have tried to find the source of consciousness starting from the brain down to the genetic code. But even at that level, the trace of consciousness is not found, though its existence can be experienced by every individual. The composition of consciousness does not seem to be purely mechanical or chemical, as discussed in section 5. We cannot even understand our mind and intelligence completely, what to speak of understanding our consciousness. Had it been composed of matter only, it could be simulated by networks of electronic circuits or by some other engineering means. If, however, it is not just matter, but something beyond matter, which it seems it is, then there is no hope for artificial consciousness.

No matter how hard we try, perhaps consciousness will always remain transcendental to human knowledge. The very source of Logic is consciousness itself. Thus, it is impossible to understand consciousness by applying Logic. How can anybody understand the source by a product of the source? Maybe one can understand to some extent, but not completely.

9. Conclusion

We have informally analyzed the foundations and frontiers of artificial intelligence and its limitations. We have discussed four components of a so-called intelligent living being, namely: body, mind, intelligence, and consciousness. The body is just gross tangible matter. But when it comes to the other three components, they are not just chemical or mechanical systems made of matter. Hardware can simulate the body in some form or other. Software can attempt to simulate the mind and intelligence with the help of tools like Logic, but the simulation will always be incomplete due to some inherent limitations. The best we can do is to provide better and better approximations, though the best approximation may lag far behind the ideal target. Going further in the hierarchy, when it comes to consciousness, the subtlest of all components, then we hit a brick wall and there is no hope.

Still, the research in AI has its own significance. Though the original goal of AI was to create thinking machines and the research towards that goal has created completely different kinds of systems far from the goal, these systems have been and will continue to be successfully applied to solve many practical problems for the benefits of the human society.

Acknowledgement

I express my deepest gratitude to Prof. Dan Willard, Department of Computer Science, SUNY, Albany, and my friend Ramnath Subramanian, Management Consultant, Deloitte & Touche. Discussions with both of them helped me develop many ideas behind this paper.

References

[1] Russel S. and Norvig P., *Artificial Intelligence: A Modern Approach*, Prentice Hall, Second Edition, 2002.

[2] Turing A., “Computing Machinery and Intelligence,” *Mind: A Quarterly Review of Psychology and Philosophy*, Vol. 59, No. 236, October 1950, pp 433-460.

[3] Gödel K.,“Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme,” *I. Monatshefte für Mathematik und Physik*, 38 (1931), pp. 173-198. Translated in English by van Heijenoort: *From Frege to Gödel*. Harvard University Press, 1971.

[4] Moravec H., *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, Cambridge, Massachusetts, 1988.

[5] *Bhagavad Gita* (*Songs of the Absolute*: an ancient Vedic text), English Translation by A. C. Bhaktivedanta Swami Prabhupada, Bhaktivedanta Book Trust, 1997, pp 210 (verse 3.42).