ARTIFICIAL INTELLIGENCE: CAN MACHINES THINK?

Animal Behavior, Rational Souls, and Clever Robots

I see these human beings walking about, interacting with each other and with myself: How do I know that they aren’t just cleverly built robots? Is there a test that would allow us always to know when we are confronted with a real “person” — a Cartesian thinking thing — instead of some programmed machine?

Descartes’ metaphysical dualism implies that the human body, being made up entirely of matter, is just a complicated machine — divinely crafted, of course, but nonetheless a machine following mechanical laws. The human mind or soul inhabits this machine, and stands (in some mysterious way) in interaction with it, such that the mind “controls” at least some of what the machine does. Similarly, things that happen within or to the machine are often consciously experienced by the mind.

Descartes also believed that non-human animals (“brutes”) were simply machines, and nothing more. He believed this on the basis of **two tests** that he describes in his *Discourse on Method* (1637). The ability to speak was Descartes’ first test. He claimed that the absence of brute speech is not due to lack of speech organs (after all, magpies and parrots can imitate the human voice) — and even if they did lack these organs, we find that deaf and dumb human beings still create a language, unlike brutes. Further, human speech is more than mere “expression of passion,” which is all that brutes are capable of performing. We must not suppose that brutes possess some “unknown language,” Descartes argues, for if this were so, then they could communicate their thoughts to us as easily as they can to each other, and they clearly do not communicate their thoughts to us.

Descartes’ second test is actually best viewed as his principle criterion, with speech being just an example. This test concerns the universality or adaptability in one’s behavior. “Reason is a universal instrument,” and thus can adapt to any contingency — for instance, developing novel strings of words for novel situations. Descartes found that various animals were exceptionally skilled at a few things — even out-performing human beings, just as an adding machine can add sums more quickly than we can. But while quite good at one or two skills, they perform horribly overall, since they are unable to adapt to the peculiarities of each new situation. (This is all quite false, of course, as the animal studies of the past century have shown; but such were Descartes’ beliefs.)

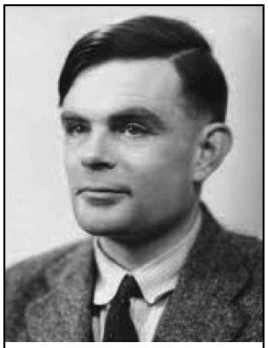
The implications of Descartes’ arguments are fairly severe. If non-human animals fail these tests, then they are understood to lack souls; and if they lack souls, then they lack mental lives, and so are fundamentally no different than human built machines, like clocks or calculators. They cannot think, nor can they suffer.

At least two questions confront us here: (1) Are these tests a proper indication of the presence of a rational mind? and (2) Can non-human animals truly not pass them? These tests were questioned from the very start, and some of Descartes’ contemporaries turned his argument in the opposite direction: Because animal behavior did not seem all that different from what humans do, if all animal behavior could be understood mechanistically, then so too could all human behavior — and thus we should think of ourselves as nothing more than machines. The most famous proponent of this view was the French philosopher and physician **Julien Offray de La Mettrie** (1709-1751) and his notorious book, *Man a Machine*.[1] Drawing a clear line between human beings and other animals has not been easy, and it is constantly being redrawn as we increase our understanding of other animals. We once thought that only humans could use tools, or could pass down information from one generation to the next, or engage in play, or deceive others, or form concepts, or have a “theory of mind” (a sense of the intentions of another individual). Each of these lines was eventually erased by ethologists and comparative psychologists who study the behavior of other animals.

Julien Offray de La Mettrie
(France, 1709-1751)

As it turns out, there actually are two lines to draw, not one — although this has not always been clear in the history of the discussion. First, we are looking for an essential difference between human beings and other animals; second, we are looking for an essential difference between human beings and humanly-built computers and robots. These are potentially quite different borders to negotiate, and I would like now to turn exclusively to a consideration of the latter border.

[1] Julien Offray de LaMettrie, *L’homme machine* (Leyden, 1748).



Alan Turing
(Great Britain, 1912-1954)

Alan Mathison Turing (1912-1954) was an English mathematician, logician, and early theorist of computer science who, among other things, built a computer used to crack the German military code (devised by their own "Enigma" machine) during World War II.

Turing was also interested in the field that is now called "artificial intelligence," and he developed the famous **Turing Test** as a criterion for deciding whether computers can indeed think.[2] This test was actually quite simple: it involved two humans, A and B, and a computer, C. The first human, A, would communicate, by way of a keyboard, with B and C. A would ask any question he liked of his two interlocutors, and if he was unable to reliably say which was the human and which the computer, then the computer was said to have "passed the test" and, for all practical purposes, would be said to be in possession of a mind (i.e., be able to *think*). It is with the articulation of this test that the field of artificial intelligence officially began.

TURING MACHINES

If it's in state:	**1**			**2**			**3**			**4**		
and it reads:	A	-	B	A	-	B	A	-	B	A	-	B
then it writes:	A	-	B	B	-	B	A	-	B	A	-	A
and moves:	r	-	r	-	-	r	r	r	l	r	-	-
and enters state:	1	1	2	3	2	2	4	4	3	1	4	4

Turing machines are the basis of all computers that exist today. The hardware to be used is left unspecified; a Turing machine could be implemented in a structure made of banana peels and egg shells, although perhaps with some difficulty. Normally, silicon chips are used to implement them. They are characterized as having a finite number of states, where a **state** is a disposition to act. The possible actions are to read a symbol, erase and/or write a symbol, move to an adjacent cell (either left or right) to read another symbol, and change to a different state. The **symbols** could be thought of as existing on a long tape, but they could just as easily be embodied in a number of different media, such as iron oxide dust on a floppy computer disk or pits in the surface of a DVD. Depending on the state that the machine is in and the symbol that is being read, the machine will perform any of the following **actions**: (i) move to the previous or next symbol, or continue reading the same symbol; (ii) erase the symbol and write another symbol; and (iii) change to a different state, or remain in the same state. The sample machine in the accompanying box is designed to take any string of A's and B's (our sample symbols) and re-order them so that all the A's come first, followed by all the B's. It's a simple machine (much simpler than one designed to add or subtract numbers), but it does its job transparently and well. It consists of four different states, which are described in terms of how the machine responds when it reads a certain symbol (A, B, or no symbol). Imagine a sample tape with the letters 'BABA', and now imagine moving between the four states of the machine, as described in this table, as you grind through the letters of the sample tape (begin in state 1 reading the 'B' on the far left). After fifteen or so moves, the sequence 'BABA' will be re-ordered as 'AABB' and the machine will stop.

MACHINE STATES AND STATES OF MIND

The view that the mind is just a fancy Turing machine is rather compelling. The states of Turing machines can be thought of as "dispositions to behave" just as minds have dispositions. If a Turing machine is in state #1, for instance, and it sees a "0", then it might erase the "0" and write a "1", move to the next symbol, and enter state #2. If I am in a hungry state and I see a pizza, then I might move to the pizza, consume a portion of it, and enter the state of satiation.

Artificial intelligence (AI) is the attempt to simulate human intelligence in a computer. It assumes a functionalist account of the mind — the mind is just the functional description of the body, primarily the brain. Therefore this function might, in theory, be replicated or modeled in a computer (thus producing artificial intelligence).

If a task can be done on a Turing machine, then that task is **algorithmic** (or computable). This is "Turing's Thesis," and was the first precise definition of what an algorithm is. A task is algorithmic, in other words, if the process for performing the task is so well defined that a mere machine can do it. It is hard to know whether a task is algorithmic until you attempt to program it onto a computer. For our purposes, the question is whether everything that the mind does is also algorithmic; if it is, then we should be able to implement or model the mind in a computer. At that point, it *might* be legitimate to say that the computer can think.

[2] Alan Turing, "Computing Machinery and Intelligence" in *Mind* 59 (1950): 433-60.

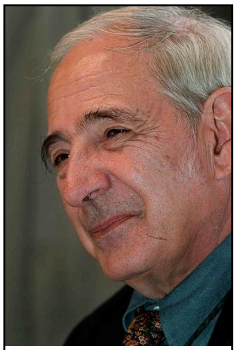
Artificial Intelligence as a “Top-Down” Strategy

One can try to explain what the mind is in either of two general ways: from the bottom-up or from the top-down. Bottom-up strategies begin with the “atoms” of mental experience and work upwards until reaching the complex phenomena of various mental skills (such as remembering, learning, and pattern-recognition). The two likeliest candidates of this bottom-up strategy are behaviorism (focusing on stimuli and responses) and a neuro-physiological approach that looks at firing patterns of individual neurons. Each of these comes with its problems: the stimuli and responses that behaviorism acknowledges aren’t likely to be the relevant atoms, and with neurophysiology, there are so many neural connections that, even while these are likely our best candidate for the “mental atoms,” the technical difficulties surrounding their exhaustive study appear to be, at least at present, insurmountable. These problems make top-down strategies more attractive. With this top-down approach, you analyze complex mental phenomena into ever smaller units of organization until you arrive at non-conscious elements (such as neurons and their connections). This strategy best characterizes AI and traditional epistemology — for instance, the most general top-down approach is Kant’s: How could anything experience or know anything?

One general strategy in AI is to analyze our mental functions into simpler and simpler functions until finally the functions, when viewed by themselves, no longer appear to be minded or intelligent. Consider the problem of how we form a visual representation of the world. A naïve view of this process, put as crudely as possible, assumes that there is a person inside your brain that interprets the images coming in, as though there were a movie screen inside the head (these are the internal representations), as well as a little person (or *homunculus*) watching the show (that is, interpreting these representations). This account, however, does little to explain how we understand the world; it just puts the problem off a step, for either the homunculus understands what he sees or he does not; if he does not, then neither do we; if he does, then there must be an even smaller homunculus inside of him, observing its own set of internal representations (and here, of course, we enter an infinite regress). Representations cannot simply understand themselves; there must be an interpreter. The approach of AI is to solve this problem by breaking down this interpreter-function into sets or structures of functions that are so simple that they do, in fact, understand themselves. The mind, as we know it, disappears into its non-mental parts, becoming nothing more than the sum-total of these parts insofar as they are functioning together.[3]

SEARLE’S CRITICISMS OF ARTIFICIAL INTELLIGENCE

John Searle (b. 1932) teaches philosophy at the University of California/Berkeley and has become a prominent critic of functionalism and the AI project. In his essay, “The Myth of the Computer” (1982), Searle notes that there are three levels for explaining human behavior. The first level is what has come to be called “**Folk psychology**,” the common-sense understanding of conscious intelligence. This consists of hundreds of common-sense generalizations or laws like “Persons in pain tend to want to relieve that pain” or “Persons who are angry tend to be impatient.” These laws make use of various concepts like belief, desire, fear, and pain, and we use these laws and concepts to explain and predict human behavior. This level of explanation works well enough in practice, but is not scientific.


John Searle
(USA, 1932-)

In the past several centuries, Searle notes, we have become convinced that our folk psychology is somehow grounded in the workings of the brain. Neurophysiology — a second level for explaining human behavior — is scientific, but not well developed, and (perhaps merely as a consequence of its immature state) it cannot explain much of our behavior.

Cognitive science is the most recent attempt at a third level between these two — a kind of a scientific psychology that is not introspective, and yet not merely a study of the brain.

Many cognitive scientists see at the heart of their field a theory of mind based on artificial intelligence, that Searle summarizes with three propositions: (1) the mind is a program, (2) the neurophysiology of the brain is irrelevant, and (3) the Turing test is the criterion of the mental. Searle criticizes each of these propositions. Against the claim that **the mind is a program**, Searle notes that the mind does one thing that no program does: it attaches an interpretation to the symbols used. As Searle puts it, computer programs are mere **syntax without semantics**; the symbols remain uninterpreted in the computer. Searle supports his criticism with what has become a famous thought-experiment: **the Chinese Room**. He asks us to imagine

[3] Cf. William Lycan’s “homuncular functionalism” as discussed in his “Form, Function, and Feel,” *Journal of Philosophy*, 78 (1981) 24-49.

a room without windows, but with something like two mail slots — one for incoming pieces of paper, and one for outgoing — and hundreds of books lining the walls inside the room. The room also contains one non-Chinese speaking human adult — call her Betty. The pieces of paper sent into the room contain sentences written in Chinese, and the books are filled with transformation rules that tell Betty how to respond (also in Chinese) to these sentences. Betty need not know that the sentences are in Chinese, or even that they are sentences. All she needs to do is identify the string of symbols in one of the books and then copy out the corresponding set of symbols that the book indicates. Now suppose that a Chinese speaker, Wenje, is writing down messages and sending them into the room, and that appropriate responses are coming back out. It would appear that Wenje is having a conversation with Betty. But by hypothesis, Betty doesn't know that the symbols she is manipulating are sentences, much less Chinese sentences, and she has no idea that she is conversing with someone. But this is precisely the situation of a computer: It shuffles symbols around following pre-set rules (the syntax), with no understanding (the interpretation or semantic content of the symbols) of the symbols. Therefore, the computer has no semantics, no understanding of the symbols.

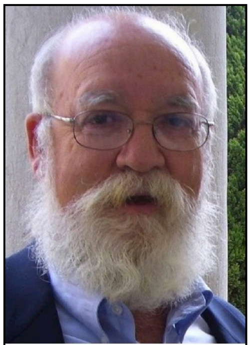
The second proposition — that **the neurophysiology of the brain is irrelevant** — seems to rest on the notion that a computer simulation is the same thing as whatever is being simulated. If we can manage to simulate the workings of the brain on a computer, then there is nothing significantly different between the two. But Searle finds this absurd. A computer might simulate the various mechanisms involved in our feeling thirsty, and even have it print out the words: "I'm thirsty" — but no one would contend that the computer really is thirsty. Much of our behavior, Searle continues, is grounded in the kind of physical beings that we are, not simply in the way that these beings function.

Searle is being tendentious here. His examples seem crazy, because computers aren't the sort of things that eat or drink (and thus are not the sort of things that get thirsty or hungry). But strong AI doesn't claim that computers are beings capable of thirst or hunger; rather, it claims that they are capable of *thought*. Thirst needs a body, but does *thinking* need a brain? Strong AI does not think so; but Searle disagrees:

> I believe that everything we have learned about human and animal biology suggests that what we call "mental" phenomena are as much a part of our biological natural history as any other biological phenomena... Much of the implausibility of the strong AI thesis derives from its resolute opposition to biology.

Finally, Searle believes that his Chinese room thought-experiment undermines **the Turing test**. Wenje, the native Chinese speaker, might easily believe that he is having a conversation with someone who understands Chinese, when by definition he is not.

Searle's arguments against AI have not gone unchallenged. **Daniel Dennett** (b. 1942) and others have argued that the Chinese Room argument fails to undermine AI because it mistakes the level at which "understanding" takes place. In the Chinese Room, Betty clearly has no understanding of Chinese, or even what she is doing — that's true by the very terms of the argument. But Dennett wishes to argue that the room itself understands Chinese. This is the "systems reply" to Searle — a reply that Searle finds preposterous. When put in terms of the thought-experiment, the systems reply might indeed seem preposterous, but Dennett would argue that this preposterousness is only an illusion caused by the terms of the argument. After all, we have entities who are clearly conscious beings — Betty, Wenje — and it's also clear that Betty understands none of the Chinese being spoken, whereas Wenje does. Because they are both (*ex hypothesi*) human beings, then it would seem that they are at the same epistemic level — but of course they are not. The entire Chinese Room is at the same level as Wenje, and inside Wenje we could postulate some analogous Betty who is equally oblivious to what is going on.


Daniel Dennett
(USA, 1942-)

What do you think?