What Your Computer Can't Know

John R. Searle

OCTOBER 9, 2014 ISSUE

The 4th Revolution: How the Infosphere Is Reshaping Human Reality
by Luciano Floridi
Oxford University Press, 248 pp., $27.95

Superintelligence: Paths, Dangers, Strategies
by Nick Bostrom
Oxford University Press, 328 pp., $29.95

Private Collection/Art Resource © 2014 C. Herscovici/Artists Rights Society (ARS), New York

René Magritte: Birth of the Idol*, 1926*

We are all beneficiaries of the revolution in computation and information technology—for example, I write this review using devices unimaginable when I was an undergraduate—but there remain enormous philosophical confusions about the correct interpretation of the technology. For example, one routinely reads that in exactly the same sense in which Garry Kasparov played and beat Anatoly Karpov in chess, the computer called Deep Blue played and beat Kasparov.

It should be obvious that this claim is suspect. In order for Kasparov to play and win, he has to be conscious that he is playing chess, and conscious of a thousand other things such as that he opened with pawn to K4 and that his queen is threatened by the knight. Deep Blue is conscious of none of these things because it is not conscious of anything at all. Why is consciousness so important? You cannot literally play chess or do much of anything else cognitive if you are totally disassociated from consciousness.

I am going to argue that both of the books under review are mistaken about the relations between consciousness, computation, information, cognition, and lots of other phenomena. So at the beginning, let me state their theses as strongly as I can. Luciano Floridi's book, *The 4th Revolution*, is essentially a work of metaphysics: he claims that in its ultimate nature, reality consists of information. We all live in the "infosphere," and we are all "inforgs" (information organisms). He summarizes his view as follows:

> *Minimally*, infosphere denotes the whole informational environment constituted by all informational entities, their properties, interactions, processes, and mutual relations…. *Maximally*, infosphere is a concept that can also be used as synonymous with reality, once we interpret the latter informationally. In this case, the suggestion is that what is real is informational and what is informational is real.

Nick Bostrom's book, *Superintelligence*, warns of the impending apocalypse. We will soon have intelligent computers, computers as intelligent as we are, and they will be followed by superintelligent computers vastly more intelligent that are quite likely to rise up and destroy us all. "This," he tells us, "is quite possibly the most important and most daunting challenge humanity has ever faced."

Floridi is announcing a completely new era. He sees himself as the successor to Copernicus, Darwin, and Freud, each of whom announced a revolution that transformed our self-conception into something more modest. Copernicus taught that we are not the center of the universe, Darwin that we are not God's special creation, Freud that we are not even masters of our own minds, and Floridi that we are not the champions of information. He claims that the revolution in ICTs (information and communication technologies) shows that everything is information and that computers are much better at it.

While Floridi celebrates the revolution, Bostrom is apocalyptic about the future. His subtitle, *Paths, Dangers, Strategies*, tells the story. There are "paths" to a superintelligent computer, the "danger" is the end of everything, and there are some "strategies," not very promising, for trying to avoid the apocalypse. Each book thus exemplifies a familiar genre: the celebration of recent progress (Floridi) and the warning of the coming disaster together with plans for avoiding it (Bostrom). Neither book is modest.

1.

The Objective-Subjective Distinction and Observer Relativity

The distinction between objectivity and subjectivity looms very large in our intellectual culture but there is a systematic ambiguity in these notions that has existed for centuries and has done enormous harm. There is an ambiguous distinction between an epistemic sense ("epistemic" means having to do with knowledge) and an ontological sense ("ontological" means having to do with existence). In the epistemic sense, the distinction is between types of claims (beliefs, assertions, assumptions, etc.). If I say that Rembrandt lived in Amsterdam, that statement is epistemically objective. You can ascertain its truth as a matter of objective fact. If I say that Rembrandt was the greatest Dutch painter that ever lived, that is evidently a matter of subjective opinion: it is epistemically subjective.

Underlying this epistemological distinction between types of claims is an ontological distinction between modes of existence. Some entities have an existence that does not depend on being experienced (mountains, molecules, and tectonic plates are good examples). Some entities exist only insofar as they are experienced (pains, tickles, and itches are examples). This distinction is between the ontologically objective and the ontologically subjective. No matter how many machines may register an itch, it is not really an itch until somebody consciously feels it: it is ontologically subjective.

A related distinction is between those features of reality that exist regardless of what we think and those whose very existence depends on our attitudes. The first class I call *observer independent* or *original*, *intrinsic*, or *absolute*. This class includes mountains, molecules, and tectonic plates. They have an existence that is wholly independent of anybody's attitude, whereas money, property, government, and marriage exist only insofar as people have certain attitudes toward them. Their existence I call *observer dependent* or *observer relative*.

These distinctions are important for several reasons. Most elements of human civilization—money, property, government, universities, and *The New York Review* to name a few examples–are observer relative in their ontology because they are created by consciousness. But the consciousness that creates them is not observer relative. It is intrinsic and many statements about these elements of civilization can be epistemically objective. For example, it is an objective fact that the *NYR* exists.

In this discussion, these distinctions are crucial because just about all of the central notions —computation, information, cognition, thinking, memory, rationality, learning, intelligence, decision-making, motivation, etc.—have two different senses. They have a sense in which they refer to actual, psychologically real, observer-independent phenomena, such as, for example, my conscious thought that the congressional elections are a few weeks away. But they also have a sense in which they refer to observer-relative phenomena, phenomena that only exist relative to certain attitudes, such as, for example, a sentence in the newspaper that says the elections are a few weeks away.

Bostrom's book is largely about computation and Floridi's book is about information. Both notions need clarification.

2.

Computation

In 1950, Alan Turing published an article in which he set out the Turing Test.[1] The purpose of the test was to establish whether a computer had genuine intelligence: if an expert cannot distinguish between human intelligent performance and computer performance, then the computer has genuine human intelligence. It is important to note that Turing called his article "Computing Machinery and Intelligence." In those days "computer" meant a person who computes. A computer was like a runner or a singer, someone who does the activity in question. The machines were not called "computers" but "computing machinery."

The invention of machines that can do what human computers did has led to a change in the vocabulary. Most of us now think of "computer" as naming a type of machinery and not as a type of person. But it is important to see that in the literal, real, observer-independent sense in which humans compute, mechanical computers do not compute. They go through a set of transitions in electronic states that we can interpret computationally. The transitions in those electronic states are absolute or observer independent, but *the computation is observer relative.* The transitions in physical states are just electrical sequences unless some conscious agent can give them a computational interpretation.

This is an important point for understanding the significance of the computer revolution. When I, a human computer, add 2 + 2 to get 4, that computation is observer independent, intrinsic, original, and real. When my pocket calculator, a mechanical computer, does the same computation, the computation is observer

relative, derivative, and dependent on human interpretation. There is no psychological reality at all to what is happening in the pocket calculator.

What then is computation? In its original, observer-independent meaning, when someone computed something, he or she figured out an answer to a question, typically a question in arithmetic and mathematics, but not necessarily. So, for example, by means of computation, we figure out the distance from the earth to the moon. When the computation can be performed in a way that guarantees the right answer in a finite number of steps, that method is called an "algorithm." But a revolutionary change took place when Alan Turing invented the idea of a Turing machine. Turing machines perform computations by manipulating just two types of symbols, usually thought of as zeroes and ones but any symbols will do. Anything other computers can do, you can do with a Turing machine.

It is important to say immediately that a Turing machine is not an actual type of machine you can buy in a store, it is a purely abstract theoretical notion. All the same, for practical purposes, the computer you buy in a store is a Turing machine. It manipulates symbols according to computational rules and thus implements algorithms.

There are two important consequences of this brief discussion, and much bad philosophy, not to mention psychology and cognitive science, has been based on a failure to appreciate these consequences.

First, *a digital computer is a syntactical machine*. It manipulates symbols and does nothing else. For this reason, the project of creating human intelligence by designing a computer program that will pass the Turing Test, the project I baptized years ago as Strong Artificial Intelligence (Strong AI), is doomed from the start. The appropriately programmed computer has a syntax but no semantics.

Minds, on the other hand, have mental or semantic content. I illustrated that in these pages with what came to be known as the Chinese Room Argument.[2] Imagine someone who doesn't know Chinese—me, for example—following a computer program for answering questions in Chinese. We can suppose that I pass the Turing Test because, following the program, I give the correct answers to the questions in Chinese, but all the same, I do not understand a word of Chinese. And if I do not understand Chinese on the basis of implementing the computer program, neither does any other digital computer solely on that basis.

This result is well known, but a second result that is just as important is made explicit in this article. Except for the cases of computations carried out by conscious human beings, *computation, as defined by Alan Turing and as implemented in actual pieces of machinery, is observer relative*. The brute physical state transitions in a piece of electronic machinery are only computations relative to some actual or possible consciousness that can interpret the processes computationally. It is an epistemically objective fact that I am writing this in a Word program, but a Word program, though implemented electronically, is not an electrical phenomenon; it exists only relative to an observer. Both of the books under review neglect these points.

3.

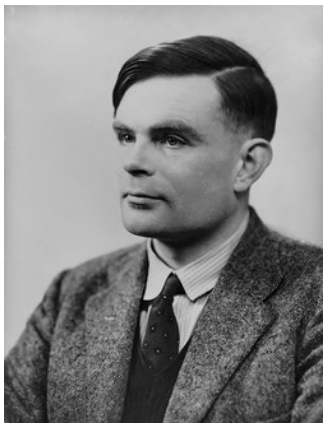
Superintelligent Computers

The picture that Bostrom has is this: we are now getting very close to the period when we will have "intelligent computers" that are as intelligent as human beings. But very soon we are almost certain to have "superintelligent computers" that are vastly more intelligent than human beings. When that happens, we are in a very serious, indeed apocalyptic, danger. The superintelligent computers might decide, on the basis of their arbitrarily formed motivations, to destroy us all—and might destroy not just their creators but all life on earth. This is for Bostrom a real threat, and he is anxious that we should face it squarely and take possible steps to prevent the worst-case scenario.

What should we say about this conception? If my account so far has been at all accurate, the conception is incoherent. If we ask, "How much real, observer-independent intelligence do computers have, whether 'intelligent' or 'superintelligent'?" the answer is zero, absolutely nothing. The intelligence is entirely observer relative. And what goes for intelligence goes for thinking, remembering, deciding, desiring, reasoning, motivation, learning, and information processing, not to mention playing chess and answering the factual questions posed on *Jeopardy!* In the observer-independent sense, the amount that the computer possesses of each of these is zero. Commercial computers are complicated electronic circuits that we have designed for certain jobs. And while some of them do their jobs superbly, do not for a moment think that there is any psychological reality to them.

Why is it so important that the system be capable of consciousness? Why isn't appropriate behavior enough? Of course for many purposes it *is* enough. If the computer can fly airplanes, drive cars, and win

at chess, who cares if it is totally nonconscious? But if we are worried about a maliciously motivated superintelligence destroying us, then it is important that the malicious motivation should be real. Without consciousness, there is no possibility of its being real.



National Portrait Gallery, London

Alan Turing, 1951

What is the argument that without consciousness there is no psychological reality to the facts attributed to the computer by the observer-relative sense of the psychological words? After all, most of our mental states are unconscious most of the time, and why should it be any different in the computer? For example, I believe that Washington was the first president even when I am sound asleep and not thinking about it. We have to distinguish between the unconscious and the nonconscious. There are all sorts of neuron firings going on in my brain that are not unconscious, they are nonconscious. For example, whenever I see anything there are neuronal feedbacks between V1 (Visual Area 1) and the LGN (lateral geniculate nucleus). But the transactions between V1 and the LGN are not unconscious mental phenomena, they are nonconscious neurobiological phenomena.

The problem with the commercial computer is it is totally nonconscious. In earlier writings,[3] I have developed an argument to show that we understand mental predicates—i.e., what is affirmed or denied about the subject of a proposition—conscious or unconscious, only so far as they are accessible to consciousness. But for present purposes, there is a simpler way to see the point. Ask yourself what fact corresponds to the claims about the psychology in both the computer and the conscious agent. Contrast my conscious thought processes in, for example, correcting my spelling and the computer's spell-check. I have a "desire" to spell correctly, and I "believe" I can find the correct spelling of a word by looking it up in a dictionary, and so I do "look up" the correct spelling. That describes the psychological reality of practical reasoning. There are three levels of description in my rational behavior: a neurobiological level, a mental or conscious level that is caused by and realized in the neurobiological level, and a level of intentional behavior caused by the psychological level.

Now consider the computer. If I misspell a word, the computer will highlight it in red and even propose alternative spellings. What psychological facts correspond to these claims? Does the computer "desire" to produce accurate spelling? And does it "believe" that I have misspelled? There are no such psychological facts at all. The computer has a list of words, and if what I type is not on the list, it highlights it in red. In the case of the computer, there are only two levels: there is the level of the hardware and the level of the behavior, but no intermediate level that is psychologically real.

Bostrom tells us that AI motivation need not be like human motivation. But all the same, there has to be some motivation if we are to think of it as engaging in motivated behavior. And so far, no sense has been given to attributing any observer-independent motivation at all to the computer.

This is why the prospect of superintelligent computers rising up and killing us, all by themselves, is not a real danger. Such entities have, literally speaking, no intelligence, no motivation, no autonomy, and no agency. We design them to behave as if they had certain sorts of psychology, but there is no psychological reality to the corresponding processes or behavior.

It is easy to imagine robots being programmed by a conscious mind to kill every recognizable human in sight. But the idea of superintelligent computers intentionally setting out on their own to destroy us, based on their own beliefs and desires and other motivations, is unrealistic because the machinery has no beliefs, desires, and motivations.

One of the strangest chapters in Bostrom's book is one on how we might produce intelligent computers by "emulating" the brain on a computer. The idea is that we would emulate each neuron as a computational device. But the computational emulation of the brain is like a computational emulation of the stomach: we could do a perfect emulation of the stomach cell by cell, but such emulations produce models or pictures and not the real thing. Scientists have made artificial hearts that work but they do not produce them by computer simulation; they may one day produce an artificial stomach, but this too would not be such an emulation.

Even with a perfect computer emulation of the stomach, you cannot then stuff a pizza into the computer and expect the computer to digest it. Cell-by-cell computer emulation of the stomach is to real digestive processes as cell-by-cell emulation of the brain is to real cognitive processes. But do not mistake the simulation (or emulation) for the real thing. It would be helpful to those trying to construct the real thing but far from an actual stomach. There is nothing in Bostrom's book to suggest he recognizes that the brain is an organ like any other, and that cells in the brain function like cells in the rest of the body on causal

biological principles.

4.

Information

Floridi does not offer a definition of information, but there is now so much literature on the subject, including some written by him, that we can give a reasonably accurate characterization. There are two senses of "information" that have emerged fairly clearly. First, there is the commonsense notion of information in which it always involves some semantic representation. So, for example, I know the way to San Jose, and that implies that I have information about how to get to San Jose. If we contrast real information with misinformation, then information, so defined, always implies truth. There is another sense of "information" that has grown up in mathematical information theory that is entirely concerned with bits of data that do not have any semantic content. But for purposes of discussing Floridi, we can concentrate on the commonsense notion because when he says that "what is real is informational and what is informational is real," he is not relying on the technical mathematical notion.

There is an immediate problem, or rather set of problems, with the idea that everything in the universe is information. First, we need to distinguish observer-independent information from observer-relative information. I really do have in my brain information about how to get to San Jose, and that information is totally observer independent. I have that regardless of what anybody thinks. The map in my car and the GPS on the dashboard also contain information about the way to San Jose, but the information there is, as the reader will recognize by now, totally observer relative.

There is nothing intrinsic to the physics that contains information. The distinction between the observer-independent sense of information, in which it is psychologically real, and the observer-relative sense, in which it has no psychological reality at all, effectively undermines Floridi's concept that we are all living in the infosphere. Almost all of the information in the infosphere is observer relative. Conscious humans and animals have intrinsic information but there is no intrinsic information in maps, computers, books, or DNA, not to mention mountains, molecules, and tree stumps. The sense in which they contain information is all relative to our conscious minds. A conscious mind surveying these objects can get the information, for example, that hydrogen atoms have one electron and that the tree is eighty-seven years old. But the atoms and the tree know nothing of this; they have no information at all.

When Floridi tells us that there is now a fourth revolution—an information revolution so that we all now live in the infosphere (like the biosphere), in a sea of information—the claim contains a confusion. The other three revolutions all identify features that are observer independent. Copernicus, Darwin, and Freud all proposed theories purporting to identify actual, observer-independent facts in the world: facts about the solar system, facts about human evolution, and facts about human unconsciousness. Even for Freud, though the unconscious requires interpretation for us to understand it, he continuously supposed that it has an existence entirely independent of our interpretations.

But when we come to the information revolution, the information in question is almost entirely in our attitudes; it is observer relative. Floridi tells us that "reality" suitably interpreted consists entirely of information. But the problem with that claim is that information only exists relative to consciousness. It is either intrinsic, observer-independent information or information in a system treated by consciousness as having information. When anybody mentions information, you ought to insist on knowing the content of the information. What is the information? What is the information about? And in what does the information consist? I do not think he offers a precise and specific answer to these questions.

Floridi's book is essentially an essay in metaphysics—metaphysics that I find profoundly mistaken. According to the metaphysics that his view is opposed to, the universe consists entirely in entities we find it convenient, if not entirely accurate, to call "particles." Maybe better terms would be "points of mass energy" or "strings," but in any case we leave it to the physicists to ascertain the basic structure of the universe. Some of these particles are organized into systems, where the boundaries of the system are set by causal relations. Examples of systems are water molecules, babies, and galaxies.

On our little Earth, some of these systems made of big carbon-based molecules with lots of hydrogen, nitrogen, and oxygen have evolved into life. And some of these life forms have evolved into animals with nervous systems. And some of these animals with nervous systems have evolved consciousness and, with consciousness, the capacity to think and express thought. Once you have consciousness and thought, you have the possibility of recognizing, creating, and sustaining information. Information is entirely a derivative higher-order phenomenon, and to put it quite bluntly, only a conscious agent can have or create information.

Here is Floridi's rival picture: information is the basic structure of the universe. All the elements of the

universe, including us, are information. What we think of as matter is patterns of information. We, as humans, are just more information. What is wrong with this picture? I do not believe it can be made consistent with what we know about the universe from atomic physics and evolutionary biology. All the literal information in the universe is either intrinsic or observer relative, and both are dependent on human or animal consciousness. Consciousness is the basis of information; information is not the basis of consciousness.

For Floridi, a model of information and information processing is in computers. How does he cope with the fact that the computer is a syntactic engine? He admits it. In a strange chapter, he says that he endorses explicitly the distinction between syntax and semantics that I have made, and he points out that the computer is a syntactical engine. But if so, how is there to be any intrinsic information in the computer? Such information exists only relative to our interpretation. To put this as bluntly as I can, I think there is a huge inconsistency between this chapter, where he grants the nonintrinsic character of the semantic information in the computer and concedes that the computer is only a syntactic engine, and earlier chapters where he insists that the computer is a paradigm of actual, real information in the world. He gets the point that the syntax of the program is never sufficient for semantic information, but he does not get the point I make in this article that even the syntax is observer relative.

I agree with Floridi's account that there is a lot more information readily available in the present era than was the case previously. If you want to know the number of troops killed at Gettysburg or the number of carbon rings in serotonin, you can find the answers to these questions more or less instantly on the Internet. The idea, however, that this has produced a revolution in ontology so that we live in a universe consisting of information seems to me not a well-defined thesis because most of the information in question is observer relative.

5.

Both of these books are rich in facts and ideas. Floridi has a good chapter on privacy in the information age, and Bostrom has extensive discussions of technological issues, but I am concentrating on the central claim of each author.

I believe that neither book gives a remotely realistic appraisal of the situation we are in with computation and information. And the reason, to put it in its simplest form, is that they fail to distinguish between the real, intrinsic observer-independent phenomena corresponding to these words and the observer-relative phenomena that also correspond to these words but are created by human consciousness.

Suppose we took seriously the project of creating an artificial brain that does what real human brains do. As far as I know, neither author, nor for that matter anyone in Artificial Intelligence, has ever taken this project seriously. How should we go about it? The absolutely first step is to get clear about the distinction between a simulation or model on the one hand, and a duplication of the causal mechanisms on the other. Consider an artificial heart as an example. Computer models were useful in constructing artificial hearts, but such a model is not an actual functioning causal mechanism. The actual artificial heart has to duplicate the causal powers of real hearts to pump blood. Both the real and artificial hearts are physical pumps, unlike the computer model or simulation.

Now exactly the same distinctions apply to the brain. An artificial brain has to literally create consciousness, unlike the computer model of the brain, which only creates a simulation. So an actual artificial brain, like the artificial heart, would have to duplicate and not just simulate the real causal powers of the original. In the case of the heart, we found that you do not need muscle tissue to duplicate the causal powers. We do not now know enough about the operation of the brain to know how much of the specific biochemistry is essential for duplicating the causal powers of the original. Perhaps we can make artificial brains using completely different physical substances as we did with the heart. The point, however, is that whatever the substance is, it has to duplicate and not just simulate, emulate, or model the real causal powers of the original organ. The organ, remember, is a biological mechanism like any other, and it functions on specific causal principles.

The difficulty with carrying out the project is that we do not know how human brains create consciousness and human cognitive processes. (Nor do we know the long-term effects that electronic communication may have on the consciousness created in brains.) Until we do know such facts, we are unlikely to be able to build an artificial brain. To carry out such a project it is essential to remember that what matters are the inner mental processes, not the external behavior. If you get the processes right, the behavior will be an expression of those processes, and if you don't get the processes right, the behavior that results is irrelevant.

That is the situation we are currently in with Artificial Intelligence. Computer engineering is useful for flying airplanes, diagnosing diseases, and writing articles like this one. But the results are for the most

part irrelevant to understanding human thinking, reasoning, processing information, deciding, perceiving, etc., because the results are all observer relative and not the real thing.

The points I am making should be fairly obvious. Why are these mistakes so persistent? There are, I believe, two basic reasons. First there is a residual behaviorism in the cognitive disciplines. Its practitioners tend to think that if you can build a machine that behaves intelligently, then it really is intelligent. The Turing Test is an explicit statement of this mistake.

Secondly there is a residual dualism. Many investigators are reluctant to treat consciousness, thinking, and psychologically real information processing as ordinary biological phenomena like photosynthesis or digestion. The weird marriage of behaviorism—any system that behaves as if it had a mind really does have a mind—and dualism—the mind is not an ordinary part of the physical, biological world like digestion—has led to the confusions that badly need to be exposed.

Letters
Making Philosophy Clear October 23, 2014

---

1 Alan Turing, "Computing Machinery and Intelligence," *Mind*, October 1950. ↩

2 John R. Searle, " The Myth of the Computer," *The New York Review*, April 29, 1982. ↩

3 John R. Searle, *The Rediscovery of the Mind* (MIT Press, 1992). ↩