2

I Married a Computer

John Searle

Kurzweil's Central Argument

Moore's Law on Integrated Circuits was first formulated by Gordon Moore, former head of Intel, in the mid-Sixties. I have seen different versions of it, but the basic idea is that better chip technology will produce an exponential increase in computer power. Every two years you get twice as much computer power and capacity for the same amount of money. Anybody who, like me, buys a new computer every few years observes Moore's Law in action. Each time I buy a new computer I pay about the same amount of money as, and sometimes even less than, I paid for the last computer, but I get a much more powerful machine. And according to

---

John Searle is Mills Professor of the Philosophy of Mind at the University of California at Berkeley and author of many books including Rationality in Action *(Cambridge: MIT Press, 2001).*

Ray Kurzweil, who is himself a distinguished software engineer and inventor, "There have been about thirty-two doublings of speed and capacity since the first operating computers were built in the 1940s."

Furthermore, we can continue to project this curve of increased computing power into the indefinite future. Moore's Law itself is about chip technology, and Kurzweil tells us that this technology will reach an upper limit when we reach the theoretical possibilities of the physics of silicon in about the year 2020. But Kurzweil tells us not to worry, because we know from evolution that some other technology will take over and "pick up where Moore's Law will have left off, without missing a beat." We know this, Kurzweil assures us, from "The Law of Accelerating Returns," which is a basic attribute of the universe; indeed it is a sublaw of "The Law of Time and Chaos." These last two laws are Kurzweil's inventions.

It is fair to say that *The Age of Spiritual Machines* is an extended reflection on the implications of Moore's Law, and is a continuation of a line of argument begun in his earlier book, *The Age of Intelligent Machines*. He begins by placing the evolution of computer technology within the context of evolution in general, and he places that within the history of the universe. The book ends with a brief history of the universe, which he calls "Time Line," beginning at the Big Bang and going to 2099.

So what, according to Kurzweil and Moore's Law, does the future hold for us? We will very soon have computers that vastly exceed us in intelligence. Why does increase in computing power automatically generate increased intelligence? Because intelligence, according to Kurzweil, is a matter of getting the right formulas in the right combination and then applying them over and over, in his sense "recursively," until the problem is solved. With sheer computational brute force, he thinks, you can solve any solvable problem. It is true, Kurzweil admits, that computational brute force is not enough by itself, and ultimately you will need "the complete set of unifying formulas that underlie intelligence." But we are well on the way to discovering these formulas: "Evolution determined an answer to this problem in a few billion years. We've made a good start in a few

thousand years. We are likely to finish the job in a few more decades."

Let us suppose for the sake of argument that we soon will have computers that are more "intelligent" than we are. Then what? This is where Kurzweil's book begins to go over the edge. First off, according to him, living in this slow, wet, messy hardware of our own neurons may be sentimentally appealing, like living in an old shack with a view of the ocean, but within a very few decades, sensible people will get out of neurons and have themselves "downloaded" onto some decent hardware. How is this to be done? You will have your entire brain and nervous system scanned, and then, when you and the experts you consult have figured out the programs exactly, you reprogram an electronic circuit with your programs and database. The electronic circuit will have more "capacity, speed, and reliability" than neurons. Furthermore, when the parts wear out they permit much easier replacement than neurons do.

So that is the first step. You are no longer locked into wet, slow, messy, and above all decaying hardware; you are upgraded into the latest circuitry. But it would be no fun just to spend life as a desktop in the office, so you will need a new body. And how is that to be done? Nanotechnology, the technology of building objects atom by atom and molecule by molecule, comes to the rescue. You replace your old body atom by atom. "We will be able to reconstruct any or all of our bodily organs and systems, and do so at the cellular level. ...We will then be able to grow stronger, more capable organs by redesigning the cells that constitute them and building them with far more versatile and durable materials." Kurzweil does not tell us anything at all about what these materials might be, but they clearly will not be flesh and blood, calcium bones and nucleoproteins.

Evolution will no longer occur in organic carbon-based materials but will pass to better stuff. However, though evolution will continue, we as individuals will no longer suffer from mortality. Even if you do something stupid like get blown up, you still keep a replacement copy of your programs and database on the shelf so you can be completely reconstructed at will. Furthermore, you can change your

whole appearance and other characteristics at will, "in a split second." You can look like Marlon Brando one minute and like Marlene Dietrich the next.

In Kurzweil's vision, there is no conflict between human beings and machines, because we will all soon, within the lifetimes of most people alive today, become machines. Strictly speaking we will become software. As he puts it, "*We will be software, not hardware*" (italics his) and can inhabit whatever hardware we like best. There will not be any difference between robots and us. "What, after all, is the difference between a human who has upgraded her body and brain using new nanotechnology, and computational technologies and a robot who has gained an intelligence and sensuality surpassing her human creators?" What, indeed? Among the many advantages of this new existence is that you will be able to read any book in just a few seconds. You could read Dante's *Divine Comedy* in less time than it takes to brush your teeth.

Kurzweil recognizes that there are some puzzling features of this utopian dream. If I have my programs downloaded onto a better brain and hardware but leave my old body still alive, which one is really me? The new robot or the old pile of junk? A problem he does not face: Suppose I make a thousand or a million copies of myself. Are they all me? Who gets to vote? Who owns my house? Who is my spouse married to? Whose driver's license is it, anyhow?

What will sex life be like in this brave new world? Kurzweil offers extended, one might even say loving, accounts. His main idea is that virtual sex will be just as good as, and in many ways better than, old-fashioned sex with real bodies. In virtual sex your computer brain will be stimulated directly with the appropriate signal without the necessity of any other human body, or even your own body. Here is a typical passage:

> Virtual touch has already been introduced, but the all-enveloping, highly realistic, visual-auditory-tactile virtual environment will not be perfected until the second decade of the twenty-first century. At this point,

> virtual sex becomes a viable competitor to the real thing. Couples will be able to engage in virtual sex regardless of their physical proximity. Even when proximate, virtual sex will be better in some ways and certainly safer. Virtual sex will provide sensations that are more intense and pleasurable than conventional sex, as well as physical experiences that currently do not exist.

The section on prostitution is a little puzzling to me:

> Prostitution will be free of health risks, as will virtual sex in general. Using wireless, very-high-bandwidth communication technologies, neither sex workers nor their patrons need to leave their homes.

But why pay, if it is all an electrically generated fantasy anyway? Kurzweil seems to concede as much when he says, "Sex workers will have competition from simulated—computer generated—partners." And, he goes on, "once the simulated virtual partner is as capable, sensual, and responsive as a real human virtual partner, who's to say that the simulated virtual partner isn't a real, albeit virtual, person?"

It is important to emphasize that all of this is seriously intended. Kurzweil does not think he is writing a work of science fiction, or a parody or satire. He is making serious claims that he thinks are based on solid scientific results. He is himself a distinguished computer scientist and inventor and so can speak with some authority about current technology. One of his rhetorical strategies is to cite earlier successful predictions he has made as evidence that the current ones are likely to come true as well. Thus he predicted within a year when a computer chess machine would be able to beat the world chess champion, and he wants us to take his prediction that we will all have artificial brains within a few decades as just more of the same sort of solidly based prediction. Because he frequently cites the IBM

chess-playing computer Deep Blue as evidence of superior intelligence in the computer, it is worth examining its significance in more detail.

Deep Blue and the Chinese Room

When it was first announced that Deep Blue had beaten Gary Kasparov, the media gave it a great deal of attention, and I suspect that the attitude of the general public was that what was going on inside Deep Blue was much the same sort of thing as what was going on inside Kasparov, only Deep Blue was better at that sort of thing and was doing a better job. This reveals a total misunderstanding of computers, and the programmers, to their discredit, did nothing to remove the misunderstanding. Here is the difference: Kasparov was consciously looking at a chessboard, studying the position and trying to figure out his next move. He was also planning his overall strategy and no doubt having peripheral thoughts about earlier matches, the significance of victory and defeat, etc. We can reasonably suppose he had all sorts of unconscious thoughts along the same lines. Kasparov was, quite literally, playing chess. None of this whatever happened inside Deep Blue. Nothing remotely like it.

Here is what happened inside Deep Blue. The computer has a bunch of meaningless symbols that the programmers use to represent the positions of the pieces on the board. It has a bunch of equally meaningless symbols that the programmers use to represent options for possible moves. The computer does not know that the symbols represent chess pieces and chess moves, because it does not know anything. As far as the computer is concerned, the symbols could be used to represent baseball plays or dance steps or numbers or nothing at all.

If you are tempted to think that the computer literally understands chess, then remember that you can use a variation on the Chinese Room Argument against the chess-playing computer. Let us call it the Chess Room Argument. Imagine that a man who does not know how to play chess is locked inside a room, and there he is given a set

of, to him, meaningless symbols. Unknown to him, these represent positions on a chessboard. He looks up in a book what he is supposed to do, and he passes back more meaningless symbols. We can suppose that if the rule book, i.e., the program, is skillfully written, he will win chess games. People outside the room will say, "This man understands chess, and in fact he is a good chess player because he wins." They will be totally mistaken. The man understands nothing of chess; he is just a computer. And the point of the parable is this: If the man does not understand chess on the basis of running the chess-playing program, neither does any other computer solely on that basis.

The Chinese Room Argument shows that just carrying out the steps in a computer program is not by itself sufficient to guarantee cognition. Imagine that I, who do not know Chinese, am locked in a room with a computer program for answering written questions, put to me in Chinese, by providing Chinese symbols as answers. If properly programmed I will provide answers indistinguishable from those of native Chinese speakers, but I still do not understand Chinese. And if I don't, neither does any other computer solely on the basis of carrying out the program. See my "Minds, Brains and Programs," *Behavioral and Brain Sciences*, Vol. 3 (1980) for the first statement of this argument. See also "The Myth of the Computer," published in the *New York Review of Books*, April 29, 1982.

The ingenuity of the hardware engineers and the programmers who programmed Deep Blue was manifested in this: from the point of view of mathematical game theory, chess is a trivial game because each side has perfect information. You know how many pieces you and your opponent have and what their locations are. You can theoretically know all of your possible moves and all of your opponent's possible countermoves. It is in principle a solvable game. The interest of chess for human beings and the problem for programmers arises from what is called a combinatorial explosion. In chess at any given point there is a finite number of possible moves. Suppose I am white and I have, say, eight possible moves. For each of these moves there is a set of possible countermoves by black and to them a set of pos-

sible moves by white, and so on up exponentially. After a few levels the number of possible positions on the board is astronomical and no human being can calculate them all. Indeed, after a few more moves the numbers are so huge that no existing computer can calculate them. At most a good chess player might calculate a few hundred.

This is where Deep Blue had the advantage. Because of the increased computational power of the machinery, it could examine 200 million positions per second; so, according to the press accounts at the time, the programmers could program the machine to follow out the possibilities to twelve levels: first white, then black, then white, and so on to the twelfth power. For some positions the machine could calculate as far as forty moves ahead. Where the human player can imagine a few hundred possible positions, the computer can scan billions.

But what does it do when it has finished scanning all these positions? Here is where the programmers have to exercise some judgment. They have to design a "scoring function." The machine attaches a numerical value to each of the final positions of each of the possible paths that developed in response to each of the initial moves. So for example a situation in which I lose my queen has a low number, a position in which I take your queen has a high number. Other factors are taken into consideration in determining the number: the mobility of the pieces (how many moves are available), the position of the pawns, etc. IBM experts are very secretive about the details of their scoring function, but they claim to use about 8,000 factors. Then, once the machine has assigned a number to all the final positions, it assigns numbers to the earlier positions leading to the final positions depending on the numbers of those final positions. The machine then selects the symbol that represents the move that leads to the highest number. It is that simple and that mechanical, though it involves a lot of symbol shuffling to get there. The real competition was not between Kasparov and the machine, but between Kasparov and a team of engineers and programmers.

Kurzweil assures us that Deep Blue was actually thinking. Indeed he suggests that it was doing more thinking than Kasparov. But

what was it thinking about? Certainly not about chess, because it had no way of knowing that these symbols represent chess positions. Was it perhaps thinking about numbers? Even that is not true, because it had no way of knowing that the symbols assigned represented numerical values. The symbols in the computer mean nothing at all to the computer. They mean something to us because we have built and programmed the computer so that it can manipulate symbols in a way that is meaningful to us. In this case we are using the computer symbols to represent chess positions and chess moves.

Now, with all this in mind, what psychological or philosophical significance should we attach to Deep Blue? It is, of course, a wonderful hardware and software achievement of the engineers and the programmers, but as far as its relevance to human psychology is concerned, it seems to me of no interest whatsoever. Its relevance is similar to that of a pocket calculator for understanding human thought processes when doing arithmetic. I was frequently asked by reporters at the time of the triumph of Deep Blue if I did not think that this was somehow a blow to human dignity. I think it is nothing of the sort. Any pocket calculator can beat any human mathematician at arithmetic. Is this a blow to human dignity? No, it is rather a credit to the ingenuity of programmers and engineers. It is simply a result of the fact that we have a technology that enables us to build tools to do things that we cannot do, or cannot do as well or as fast, without the tools.

Kurzweil also predicts that the fact that a machine can beat a human being in chess will lead people to say that chess was not really important anyway. But I do not see why. Like all games, chess is built around the human brain and body and its various capacities and limitations. The fact that Deep Blue can go through a series of electrical processes that we can interpret as "beating the world champion at chess" is no more significant for human chess playing than it would be significant for human football playing if we built a steel robot which could carry the ball in a way that made it impossible for the robot to be tackled by human beings. The Deep Blue chess player is as irrelevant to human concerns as is the Deep Blue running back.

Some Conceptual Confusions

I believe that Kurzweil's book exhibits a series of conceptual confusions. These are not all Kurzweil's fault; they are common to the prevailing culture of information technology, and especially to the subculture of artificial intelligence, of which he is a part. Much of the confusion in this entire field derives from the fact that people on both sides of the debate tend to suppose that what is at issue is the success or failure of computational simulations. Are human beings "superior" to computers or are computers superior? That is not the point at issue at all. The question is not whether computers can succeed at doing this or that. For the sake of argument, I am just going to assume that everything Kurzweil says about the increase in computational power is true. I will assume that computers both can and will do everything he says they can and will do, that there is no question about the capacity of human designers and programmers to build ever faster and more powerful pieces of computational machinery. My point is that to the issues that really concern us about human consciousness and cognition, these successes are irrelevant.

What, then, is at issue? Kurzweil's book exhibits two sets of confusions, which I shall consider in order.

(1) He confuses the computer simulation of a phenomenon with a duplication or re-creation of that phenomenon. This comes out most obviously in the case of consciousness. Anybody who is seriously considering having his "program and database" downloaded onto some hardware ought to wonder whether or not the resulting hardware is going to be conscious. Kurzweil is aware of this problem, and he keeps coming back to it at various points in his book. But his attempt to solve the problem can only be said to be plaintive. He does not claim to know that machines will be conscious, but he insists that they will claim to be conscious, and will continue to engage in discussions about whether they are conscious, and consequently their claims will be largely accepted. People will eventually just come to accept without question that machines are conscious.

But this misses the point. I can already program my computer so that it says that it is conscious—i.e., it prints out "I am conscious"—and a good programmer can even program it so that it will carry on a rudimentary argument to the effect that it is conscious. But that has nothing to do with whether or not it really is conscious. Actual human brains cause consciousness by a series of specific neurobiological processes in the brain. What the computer does is a simulation of these processes, a symbolic model of the processes. But the computer simulation of brain processes that produce consciousness stands to real consciousness as the computer simulation of the stomach processes that produce digestion stands to real digestion. You do not cause digestion by doing a computer simulation of digestion. Nobody thinks that if we had the perfect computer simulation running on the computer, we could stuff a pizza into the computer and it would thereby digest it. It is the same mistake to suppose that when a computer simulates the processes of a conscious brain it is thereby conscious.

The computer, as we saw in our discussion of the chess-playing program, succeeds by manipulating formal symbols. The symbols themselves are quite meaningless; they have only the meaning we have attached to them. The computer knows nothing of this, it just shuffles the symbols. And those symbols are not by themselves sufficient to guarantee equivalent causal powers to actual biological machinery like human stomachs and human brains.

Kurzweil points out that not all computers manipulate symbols. Some recent machines simulate the brain by using networks of parallel processors called "neural nets," which try to imitate certain features of the brain. But that is no help. We know from the Church-Turing Thesis, a mathematical result, that any computation that can be carried out on a neural net can be carried out on a symbol-manipulating machine. The neural net gives no increase in computational power. And simulation is still not duplication.

But, someone is bound to ask, can you prove that the computer is not conscious? The answer to this question is: Of course not. I cannot prove that the computer is not conscious, any more than I can

prove that the chair I am sitting on is not conscious. But that is not the point. It is out of the question, for purely neurobiological reasons, to suppose that the chair or the computer is conscious. The point for the present discussion is that the computer is not designed to be conscious. It is designed to manipulate symbols in a way that carries out the steps in an algorithm. It is not designed to duplicate the actual causal powers of the brain to cause consciousness. It is designed to enable us to simulate any process that we can describe precisely.

Kurzweil is aware of this objection and tries to meet it with a slippery-slope argument: We already have brain implants, such as cochlear implants in the auditory system, that can duplicate and not merely simulate certain brain functions. What is to prevent us from a gradual replacement of all the brain anatomy that would preserve and not merely simulate our consciousness and the rest of our mental life? In answer to this, I would point out that he is now abandoning the main thesis of the book, which is that what is important for consciousness and other mental functions is entirely a matter of computation. In his words, we will become software, not hardware.

I believe that there is no objection in principle to constructing an artificial hardware system that would duplicate the powers of the brain to cause consciousness using some chemistry different from neurons. But to produce consciousness any such system would have to duplicate the actual causal powers of the brain. And we know, from the Chinese Room Argument, that computation by itself is insufficient to guarantee any such causal powers, because computation is defined entirely in terms of the manipulation of abstract formal symbols.

(2) The confusion between simulation and duplication is a symptom of an even deeper confusion in Kurzweil's book, and that is between those features of the world that exist intrinsically, or independently of human observation and conscious attitudes, and those features of the world that are dependent on human attitudes—the distinction, in short, between features that are observer-independent and those that are observer-relative.

Examples of observer-independent features are the sorts of things discussed in physics and chemistry. Molecules, and mountains, and tectonic plates, as well as force, mass, and gravitational attraction, are all observer-independent. Since relativity theory we recognize that some of their limits are fixed by reference to other systems, but none of them are observer-dependent in the sense of requiring the thoughts of conscious agents for their existence. On the other hand, such features of the world as money, property, marriage, government, and football games require conscious observers and agents in order for them to exist as such. A piece of paper has intrinsic or observer-independent chemical properties, but a piece of paper is a dollar bill only in a way that is observer-dependent or observer-relative.

In Kurzweil's book many of his crucial notions oscillate between having a sense that is observer-independent, and another sense that is observer-relative. The two most important notions in the book are intelligence and computation, and both of these exhibit precisely this ambiguity. Take intelligence first.

In a psychological, observer-independent sense, I am more intelligent than my dog, because I can have certain sorts of mental processes that he cannot have, and I can use these mental capacities to solve problems that he cannot solve. But in this psychological sense of intelligence, wristwatches, pocket calculators, computers, and cars are not candidates for intelligence, because they have no mental life whatever.

In an observer-relative sense, we can indeed say that lots of machines are more intelligent than human beings because we have designed the machines in such a way as to help us solve problems that we cannot solve, or cannot solve as efficiently, in an unaided fashion. Chess-playing machines and pocket calculators are good examples. Is the chess-playing machine really more intelligent at chess than Kasparov? Is my pocket calculator more intelligent than I at arithmetic? Well, in an intrinsic or observer-independent sense, of course not, the machine has no intelligence whatever, it is just an electronic circuit that we have designed, and can ourselves operate, for certain purposes. But in the metaphorical or observer-relative

sense, it is perfectly legitimate to say that the chess-playing machine has more intelligence, because it can produce better results. And the same can be said for the pocket calculator.

There is nothing wrong with using the word "intelligence" in both senses, provided you understand the difference between the observer-relative and the observer-independent. The difficulty is that this word has been used as if it were a scientific term, with a scientifically precise meaning. Indeed, many of the exaggerated claims made on behalf of "artificial intelligence" have been based on this systematic confusion between observer-independent, psychologically relevant intelligence and metaphorical, observer-relative, psychologically irrelevant ascriptions of intelligence. There is nothing wrong with the metaphor as such; the only mistake is to think that it is a scientifically precise and unambiguous term. A better term than "artificial intelligence" would have been "simulated cognition."

Exactly the same confusion occurs over the notion of "computation." There is a literal sense in which human beings are computers because, for example, we can compute 2+2=4. But when we design a piece of machinery to carry out that computation, the computation 2+2=4 exists only relative to our assignment of a computational interpretation to the machine. Intrinsically, the machine is just an electronic circuit with very rapid changes between such things as voltage levels. The machine knows nothing about arithmetic just as it knows nothing about chess. And it knows nothing about computation either, because it knows nothing at all. We use the machinery to compute with, but that does not mean that the computation is intrinsic to the physics of the machinery. The computation is observer-relative, or to put it more traditionally, "in the eye of the beholder."

This distinction is fatal to Kurzweil's entire argument, because it rests on the assumption that the main thing humans do in their lives is compute. Hence, on his view, if—thanks to Moore's Law—we can create machines that can compute better than humans, we have equaled and surpassed humans in all that is distinctively human. But in fact humans do rather little that is literally computing. Very little of our time is spent working out algorithms to figure out answers to

questions. Some brain processes can be usefully described as if they were computational, but that is observer-relative. That is like the attribution of computation to commercial machinery, in that it requires an outside observer or interpreter.

Another result of this confusion is a failure on Kurzweil's part to appreciate the significance of current technology. He describes the use of strands of DNA to solve the Traveling Salesman Problem—the problem of how to plot a route for a salesman so that he never goes through the same city twice—as if it were the same sort of thing as the use, in some cases, of neural implants to cure Parkinson's Disease. But the two cases are completely different. The cure for Parkinson's Disease is an actual, observer-independent causal effect on the human brain. But the sense in which the DNA strands stand for or represent different cities is entirely observer-relative. The DNA knows nothing about cities.

It is worth pointing out here that when Alan Turing first invented the idea of the computer, the word "computer" meant "person who computes." "Computer" was like "runner" or "skier." But as commercial computers have become such an essential part of our lives, the word "computer" has shifted in meaning to mean "machinery designed by us to use for computing," and, for all I know, we may go through a change of meaning so that people will be said to be computers only in a metaphorical sense. It does not matter as long as you keep the conceptual distinction clear between what is intrinsically going on in the machinery, however you want to describe it, and what is going on in the conscious thought processes of human beings. Kurzweil's book fails throughout to perceive these distinctions.

The Problem of Consciousness

We are now in the midst of a technological revolution that is full of surprises. No one thirty years ago was aware that one day household computers would become as common as dishwashers. And those of us who used the old Arpanet of twenty years ago had no idea that it would evolve into the Internet. This revolution cries out for interpre-

tation and explanation. Computation and information processing are both harder to understand and more subtle and pervasive in their effects on civilization than were earlier technological revolutions such as those of the automobile and television. The two worst things that experts can do when explaining this technology to the general public are first to give the readers the impression that they understand something they do not understand, and second to give the impression that a theory has been established as true when it has not.

Kurzweil's book suffers from both of these defects. The title of the book is *The Age of Spiritual Machines*. By "spiritual," Kurzweil means conscious, and he says so explicitly. The implications are that if you read his book you will come to understand the machines and that we have overwhelming evidence that they now are or will shortly be conscious. Both of these implications are false. You will not understand computing machinery from reading Kurzweil's book. There is no sustained effort to explain what a computer is and how it works. Indeed one of the most fundamental ideas in the theory of computation, the Church-Turing Thesis, is stated in a way which is false.

Here is what Kurzweil says:

> This thesis says that all problems that a human being can solve can be reduced to a set of algorithms, supporting the idea that machine intelligence and human intelligence are essentially equivalent.

That definition is simply wrong. The actual thesis comes in different formulations (Church's is different from Turing's, for example), but the basic idea is that any problem that has an algorithmic solution can be solved on a Turing machine, a machine that manipulates only two kinds of symbols, the famous zeroes and ones.

Where consciousness is concerned, the weaknesses of the book are even more disquieting. One of its main themes, in some ways the main theme, is that increased computational power gives us good, indeed overwhelming, reason to think we are moving into an era

when computing machinery artifacts, machines made by us, will be conscious, "the age of spiritual machines." But from everything we know about the brain, and everything we know about computation, increased computational power in a machine gives us no reason whatever to suppose that the machine is duplicating the specific neurobiological powers of the brain to create consciousness. Increased computer power by itself moves us not one bit closer to creating a conscious machine. It is just irrelevant.

Suppose you took seriously the project of building a conscious machine. How would you go about it? The brain is a machine, a biological machine to be sure, but a machine all the same. So the first step is to figure out how the brain does it and then build an artificial machine that has an equally effective mechanism for causing consciousness. These are the sorts of steps by which we built an artificial heart. The problem is that we have very little idea of how the brain does it. Until we do, we are most unlikely to produce consciousness artificially in nonbiological materials. When it comes to understanding consciousness, ours is not the age of spiritual machines. It is more like the age of neurobiological infancy, and in our struggles to get a mature science of the brain, Moore's Law provides no answers.

A Brief Recapitulation

In response to my initial review of Kurzweil's book in *The New York Review of Books*, Kurzweil wrote both a letter to the editor and a more extended rebuttal on his website. He claims that I presented a "distorted caricature" of his book, but he provided no evidence of any distortion. In fact I tried very hard to be scrupulously accurate both in reporting his claims and in conveying the general tone of futuristic techno-enthusiasm that pervades the book. So at the risk of pedantry, let's recapitulate briefly the theses in his book that I found most striking:

(1) Kurzweil thinks that within a few decades we will be able to download our minds onto computer hardware. We will continue to exist as computer software. "*We will be software, not hardware*" (p. 129, his italics). And "the essence of our identity will switch to the permanence of our software" (p.129).

(2) According to him, we will be able to rebuild our bodies, cell by cell, with different and better materials using "nanotechnology." Eventually, " there won't be a clear difference between humans and robots" (p.148).

(3) We will be immortal, not only because we will be made of better materials, but because even if we were destroyed we will keep copies of our programs and databases in storage and can be reconstructed at will. "Our immortality will be a matter of being sufficiently careful to make frequent back-ups," he says, adding the further caution: "If we're careless about this, we'll have to load an old backup copy and be doomed to repeat our recent past" (p. 129). (What is this supposed to mean? That we will be doomed to repeat our recent car accident and spring vacation?)

(4) We will have overwhelming evidence that computers are conscious. Indeed there will be "no longer any clear distinction between humans and computers" (p. 280).

(5) There will be many advantages to this new existence, but one he stresses is that virtual sex will soon be a "viable competitor to the real thing," affording "sensations that are more intense and pleasurable than conventional sex" (p. 147).

Frankly, had I read this as a summary of some author's claims, I might think it must be a "distorted caricature," but Kurzweil did in fact make each of these claims, as I show by extensive quotation. In his letter he did not challenge me on any of these central points. He conceded by his silence that my understanding of him on these central issues is correct. So where is the "distorted caricature?"

I then point out that his arguments are inadequate to establish any of these spectacular conclusions. They suffer from a persistent confusion between simulating a cognitive process and duplicating it, and even worse confusion between the observer-relative, in-the-eye-of-the-beholder sense of concepts like intelligence, thinking, etc., and the observer-independent intrinsic sense.

What has he to say in response? Well, about the main argument he says nothing. About the distinction between simulation and duplication, he says he is describing neither simulations of mental powers nor re-creations of the real thing, but "functionally equivalent re-creations." But the notion "functionally equivalent" is ambiguous precisely between simulation and duplication. What exactly functions to do exactly what? Does the computer simulation function to enable the system to have *external* behavior, which is *as if* it were conscious, or does it function to actually cause *internal* conscious states? For example, my pocket calculator is "functionally equivalent" to (indeed better than) me in producing answers to arithmetic problems, but it is not thereby functionally equivalent to me in producing the conscious thought processes that go with solving arithmetic problems. Kurzweil's argument about consciousness is based on the assumption that the external behavior is overwhelming evidence for the presence of the internal conscious states. He has no answer to my objection that once you know that the computer works by shuffling symbols, its behavior is no evidence at all for consciousness. The notion of functional equivalence does not overcome the distinction between simulation and duplication; it just disguises it for one step.

In his letter he told us he is interested in doing "reverse engineering" to figure out how the brain works. But in the book there is virtually nothing about the actual working of the brain and how the specific electro-chemical properties of the thalamo-cortical system could produce consciousness. His attention rather is on the computational advantages of superior hardware.

On the subject of consciousness there actually is a "distorted caricature," but it is Kurzweil's distorted caricature of my arguments. He said, "Searle would have us believe that you can't be conscious if you don't squirt neurotransmitters (or some other specific biological process)." Here is what I actually wrote: "I believe there is no objection in principle to constructing an artificial hardware system that would duplicate the causal powers of the brain to cause consciousness using some chemistry different from neurons." Not much about the necessity of squirting neurotransmitters there. The point I made, and repeat again, is that because we know that brains cause consciousness with specific biological mechanisms, any nonbiological mechanism has to share with brains the causal power to do it. An artificial brain might succeed by using something other than carbon-based chemistry, but just shuffling symbols is not enough, by itself, to guarantee those powers. Once again, he offers no answer to this argument.

He challenges my Chinese Room Argument, but he seriously misrepresents it. The argument is not the circular claim that I do not understand Chinese because I am just a computer, but rather that I don't, as a matter of fact, understand Chinese and could not acquire an understanding by carrying out a computer program. There is nothing circular about that. His chief counterclaim is that the man is only the central processing unit, not the whole computer. But this misses the point of the argument. The reason the man does not understand Chinese is that he does not have any way to get from the symbols, the syntax, to what the symbols mean, the semantics. But if the man cannot get the semantics from the syntax alone, neither can the whole computer. It is, by the way, a misunderstanding on his part to think that I am claiming that a man could actually carry out the billions of

steps necessary to carry out a whole program. The point of the example is to illustrate the fact that the symbol manipulations alone, even billions of them, are not constitutive of meaning or thought content, conscious or unconscious. To repeat, the syntax of the implemented program is not semantics.

Concerning other points in his letter: He says that I am wrong to think that he attributes superior thinking to Deep Blue. But here is what he wrote in response to the charge that Deep Blue just does number crunching and not thinking: "One could say that the opposite is the case, that Deep Blue was indeed thinking through the implications of each move and countermove, and that it was Kasparov who did not have the time to think very much during the tournament" (p. 290).

He also says that on his view Moore's Law is only a part of the story. Quite so. In my review I mention other points he makes such as, importantly, nanotechnology.

I cannot recall reading a book in which there is such a huge gulf between the spectacular claims advanced and the weakness of the arguments given in their support. Kurzweil promises us our minds downloaded onto decent hardware, new bodies made of better stuff, evolution without DNA, better sex without the inconvenience of actual partners, computers that convince us that they are conscious, and above all personal immortality. The main theme of my critique is that the existing technological advances that are supposed to provide evidence in support of these predictions, wonderful though they are, offer no support whatever for these spectacular conclusions. In every case the arguments are based on conceptual confusions. Increased computational power by itself is no evidence whatever for consciousness in compuuers.