6

Locked in His Chinese Room: Response to John Searle

Ray Kurzweil

Those Who Build Chinese Rooms are Doomed to Live in Them

John Searle is popular among his followers for what they believe is a staunch defense of the deep mystery of human consciousness against trivialization by strong AI reductionists like Ray Kurzweil. And even though I have always found Searle's logic in his celebrated Chinese Room Argument to be hopelessly tautological, even I had expected him to articulate an elevating treatise on the paradoxes of consciousness. Thus it is with some surprise that I find Searle writing statements such as:

> [H]uman brains cause consciousness by a series of specific neurobiological processes in the brain. The essential thing is to recognize that consciousness is a biological process like digestion, lactation, photosynthesis, or mitosis . . .
>
> 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.
>
> We know that brains cause consciousness with specific biological mechanisms . . .

So who is being the reductionist here? Searle apparently expects that we can measure the subjectivity of another entity as readily as we measure the oxygen output of photosynthesis.

I will return to this central issue, but I also need to point out the disingenuous nature of many of Searle's quotations and characterizations. For example, he leaves out critical words that dramatically alter the meaning of a statement. For example, Searle writes in his chapter in this book:

> [Ray Kurzweil] insists that they [the machines] will claim to be 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.

Searle fails to point out that I make exactly the same point, and further that I refer not to such idle claims that are easily feasible today but rather to the *convincing* claims of future machines. As one example of many, I write in my book (p. 60) that these claims "won't seem like a programmed response. The machines will be earnest and convincing."

Searle writes that I "frequently cite IBM's Deep Blue as evidence of superior intelligence in the computer." The opposite is the case: I cite Deep Blue to (p. 289) "examine the human and [contemporary] machine approaches to chess . . . not to belabor the issue of chess, but rather because [they] *illustrate a clear contrast*." Human thinking follows a very different paradigm. Solutions emerge in the human brain from the unpredictable interaction of millions of simultaneous self-organizing chaotic processes. There are profound advantages to the human paradigm: we can recognize and respond to extremely subtle patterns. But we can build machines the same way.

Searle states that my book "is an extended reflection of the implications of Moore's Law." But the exponential growth of computing power is only a small part of the story. As I repeatedly state, adequate computational power is *a necessary but not sufficient condition* to achieve human levels of intelligence. Searle essentially doesn't mention my primary thesis: We are learning how to organize these increasingly formidable resources by reverse engineering the human brain itself. By examining brains in microscopic detail, we will be able to recreate and then vastly extend these processes. As I point out below, we have made substantial progress in this endeavor just in the brief period of time since my book was published.

Searle is best known for his "Chinese Room" analogy and has presented various formulations of it over twenty years (see below). His descriptions illustrate a failure to understand the essence of either brain processes or the nonbiological processes that could replicate them. Searle starts with the assumption that the "man" in the room doesn't understand anything because, after all, "he is just a computer," thereby illuminating Searle's own bias. Searle then concludes—no surprise—that the computer doesn't understand. Searle combines this tautology with a basic contradiction: The computer doesn't understand Chinese, yet (according to Searle) can convincingly answer questions in Chinese. But if an entity—biological or otherwise—really doesn't understand human language, it will quickly be unmasked by a competent interlocutor. In addition, for the program to convincingly respond, it would have to be as complex as a human brain. The observers would long be dead while the man in the room spends millions of years following a program billions of pages long.

Most importantly, the man is acting only as the central processing unit, a small part of a system. While the man may not see it, the understanding is distributed across the entire pattern of the program itself and the billions of notes he would have to make to follow the program. *I understand English, but none of my neurons do.* My understanding is represented in vast patterns of neurotransmitter strengths, synaptic clefts, and interneuronal connections. Searle appears not to understand the significance of distributed patterns of information and their emergent properties.

Searle writes that I confuse a simulation for a recreation of the real thing. What my book (and chapter in this book) actually talk about is a third category: functionally equivalent recreation. He writes that we could not stuff a pizza into a computer simulation of the stomach and expect it to be digested. But we could indeed accomplish this with a properly designed artificial stomach. I am not talking about a mere "simulation" of the human brain as Searle construes it, but rather functionally equivalent recreations of its causal powers. As I pointed out, we already have functionally equivalent replacements of portions of the brain to overcome such disabilities as deafness and Parkinson's disease.

Searle writes: "It is out of the question . . . to suppose that . . . the computer is conscious." Given this assumption, Searle's conclusions to the same effect are hardly a surprise. Searle would have us believe that you can't be conscious if you don't possess some specific (albeit unspecified) biological process. No entities based on functionally equivalent processes need apply. This biology-centric view of consciousness is likely to go the way of other human-centric beliefs. In my view, we cannot penetrate the ultimate reality of subjective experience with objective measurement, which is why many classical methods, including Searle's materialist approach, quickly hit a wall.

The Intuitive Linear View Revisited

Searle's slippery and circular arguments aside, nonbiological entities, which today have many narrowly focused skills, are going to vastly expand in the breadth, depth, and subtlety of their intelligence and creativity. Early in his chapter, Searle makes clear his discomfiture with the radical nature of the twenty-first century technologies that I have described and their impact on society. Searle clearly expects the twenty-first century to be much like the twentieth century, and considers any significant deviation from present norms to be absurd on their face. Not once, but twice he expresses incredulity at the notion of virtual sex, for example: "The section on prostitute is a little puzzling to me. . . . But why pay, if it is all an electrically generated fantasy anyway?"

Searle obviously misses the point of virtual reality. Virtual reality is not fantasy; it is a communication medium between two or more people. We already have auditory virtual reality; it's called the telephone. Indeed, that is exactly how the telephone was viewed when it was introduced in the late nineteenth century. People found it remarkable that you could actually "be with" someone else, at least as far as the auditory sense was concerned, despite the fact that you were geographically disparate. And indeed we have a form of sex over phone lines, not very satisfying to many perhaps, but keep in mind it involves only one sense organ. The paradigm, however, is just this: two people communicating, and in some cases one of those persons may be paid for their services. Technology to provide full immersion *visual* shared environments is now being developed, and will be ubiquitous by the end of this decade (with images written directly to our retinas by our eyeglasses and contact lenses). Then, in addition to talking, it will really appear like you are with that other person. As for touching one another, the tactile sense will not be full immersion by the end of this decade, but full immersion virtual shared environments incorporating the auditory, visual, and tactile senses will become available by around 2020. The design of such technology can already be described. When nanobot-based virtual reality becomes feasible around 2030, then these shared environments will encompass all of the senses.

Virtual sex and virtual prostitution are among the more straightforward scenarios for applying full immersion communication technologies, so it is puzzling to me that Searle consistently cites these as among the most puzzling to him. Clearly Searle's thinking about the future is limited by what I referred to in my chapter as the "intuitive linear" view, despite the fact that both he and I have been around long enough to witness the acceleration inherent in the historically accurate exponential view of history and the future.

Twenty-First Century Machine Intelligence Revisited

Beyond Searle's circular, tautological, and often contradictory reasoning, he essentially fails to even address the key points in my chapter and my book, so it is worthwhile reviewing my primary reasoning in my own words. My message concerns the emergence early in the next century of nonbiological entities with enormously powerful intellectual skills and abilities and the profound impact this will have on human society. The primary themes are:

> (1) The power of computer technology per unit cost is growing exponentially. This has been true for the past one hundred years, and will continue well into the next century.
>
> (2) New hardware technologies such as nanotube-based circuits, which allow three-dimensional computer circuits to be constructed, are already working in laboratories. Such three-dimensional circuits will

ultimately provide physically small devices that vastly exceed the memory and computational ability of the human brain.

(3) In addition to computation, there is comparable exponential growth in communication, brain scanning, neuron modeling, brain reverse engineering, miniaturization of technology, and many other areas.

(4) Sufficient computational power by itself is not enough. Adequate computational (and communication) resources are a necessary but not sufficient condition to match the breadth, depth, and subtlety of human capabilities. The organization, content, and embedded knowledge of these resources (i.e., the "software" of intelligence) is also critical.

(5) A key resource for understanding and ultimately recreating the software of intelligence is the human brain itself. By probing the human brain, we are already learning its methods. We are already applying these types of insights (e.g., the front-end sound-wave transformations used in automatic speech recognition systems are based on early auditory processing in mammalian brains). The brain is not invisible to us. Our ability to scan and understand human neural functioning both invasively and noninvasively is scaling up exponentially.

(6) We have already created detailed replications of substantial neuron clusters. These replications (not to be confused with the simplified mathematical models used in many contemporary "neural nets") recreate the highly parallel analog-digital functions of these neuron clusters, and such efforts are also scaling up

exponentially. This has nothing to do with manipulating symbols, but is a detailed and realistic recreation of what Searle refers to as the "causal powers" of neuron clusters. Human neurons and neuron clusters are certainly complicated, but their complexity is not beyond our ability to understand and recreate using other mediums. I cite specific recent progress below.

(7) We've already shown that the causal powers of substantial neuron clusters cannot only be recreated, but actually placed in the human brain to replace disabled brain portions. These are not mere simulations, but functionally equivalent recreations of the causal powers of neuron clusters.

(8) With continuing exponential advances in computer hardware, neuron modeling, and human brain scanning and understanding, it is a conservative statement to say that we will have detailed models of neurons and complete maps of the human brain within thirty years that enable us to reverse engineer its organization and content. This is no more startling a proposition than was the proposal to scan the entire human genome 14 years ago. Well before that, we will have nonbiological hardware with the requisite capacity to replicate its causal powers. Human brain level computational power, together with an understanding of the organization and content of human intelligence gained through such reverse engineering efforts, will be a formidable combination.

(9) Although contemporary computers can compete with human intelligence in narrow domains (e.g., chess, diagnosing blood cell images, recognizing land terrain images in a cruise missile, making financial

investment decisions), their overall intelligence lacks the subtlety and range of human intelligence. Compared to humans, today's machines appear brittle and formulaic. But contemporary computers are still a million times simpler than the human brain. The depth and breadth of the behavior of nonbiological entities will appear quite different when the million-fold difference in complexity is reversed, and when we can apply powerful models of biological processes.

(10) There are profound advantages to nonbiological intelligence. If I spend years learning French, I can't transfer that knowledge to you. You have to go through a similar painstaking process. We cannot easily transfer (from one person to another) the vast pattern of neurotransmitter strengths, synaptic clefts, and other neural elements that represents our human knowledge. But we won't leave out quick downloading ports in our nonbiological recreations of neuron clusters. Machines will be able, therefore, to rapidly share their knowledge.

(11) Virtual personalities can claim to be conscious today, but such claims are not convincing. They lack the subtle and profound behavior that would make such claims compelling. But the claims of nonbiological entities some decades from now—entities that are based on the detailed design of human thinking—will not be so easily dismissed.

(12) The emergence of highly advanced intelligence in our machines will have a profound impact on all aspects of our human-machine civilization.

Recent Progress in Brain Reverse Engineering

Critical to my thesis is the issue of brain reverse engineering, so it is worth commenting on recent progress in this area. Just in the two years since my recent book was published, progress in this area has been remarkably fast. The pace of brain reverse engineering is only slightly behind the availability of the brain scanning and neuron structure information. There are many contemporary examples, but I will cite just one, which is a comprehensive model of a significant portion of the human auditory processing system that Lloyd Watts <www.lloydwatts.com> has developed from both neurobiology studies of specific neuron types and brain interneuronal connection information. Watts' model includes more than a dozen specific brain modules, five parallel paths and includes the actual intermediate representations of auditory information at each stage of neural processing. Watts has implemented his model as real-time software which can locate and identify sounds with many of the same properties as human hearing. Although a work in progress, the model illustrates the feasibility of converting neurobiological models and brain connection data into working functionally equivalent recreations. Also, as Hans Moravec and others have speculated, these efficient machine implementations require about 1,000 times less computation than the theoretical potential of the biological neurons being recreated.

The brain is not one huge "tabula rasa" (i.e., undifferentiated blank slate), but rather an intricate and intertwined collection of hundreds of specialized regions. The process of "peeling the onion" to understand these interleaved regions is well underway. As the requisite neuron models and brain interconnection data becomes available, detailed and implementable models such as the auditory example above will be developed for all brain regions.

On the Contrast Between Deep Blue and Human Thinking

To return to Searle's conceptions and misconceptions, he misconstrues my presentation of Deep Blue. As I mentioned above, I discuss Deep Blue because it illustrates a clear contrast between this particular approach to building machines that perform certain structured tasks such as playing chess, and the way that the human brain works. In my book, I use this discussion to present a proposal to build these systems in a different way—a more human way (see below). Searle concentrates entirely on the methods used by Deep Blue, which completely misses the point.

Searle's chapter is replete with misquotations. For example, Searle states:

> 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.

This is a completely erroneous reference. I repeatedly state that increases in computing power do not automatically generate increased intelligence. Furthermore, with regard to Searle's reference to recursion, I present the recursive method as only one technique among many, and as a method suitable only for a narrow class of problems such as playing board games. I never present this simple approach as the way to create human-level intelligence in a machine.

If you read Searle's chapter and do not read my book, you would get the impression that I present the method used by Deep Blue as the ultimate paradigm for machine intelligence. It makes me wonder if Searle actually read the book, or just selectively picked phrases out of context. I repeatedly contrast the recursive methods of Deep Blue with the pattern recognition based paradigm used by the human brain. The field of pattern recognition represents my own technical area of expertise. Human pattern recognition is based on a paradigm in which solutions emerge from the interplay of many interacting processes (see below). What I clearly describe in the book is moving away from the formulaic approaches used by many contemporary AI systems and moving towards the human paradigm of pattern recognition.

Searle's explanation of how Deep Blue works is essentially correct (thanks in large measure to my explaining it to him in response to his urgent email messages to me asking me to clarify for him how Deep Blue works). Although the basic recursive method of rapidly expanding move-countermove sequences is simple, the evaluation at the "leaves" of this move-countermove tree (the scoring function) is really the heart of the method. If you have a simple scoring function, then the method is indeed simple and dependent merely on brute force in computational speed. However, the scoring function is not necessarily simple. Deep Blue's scoring function uses up to 8,000 different features, and is more complex than most.

Deep Blue is able to consider billions of board situations and creates an enormous tree of move-countermove possibilities. Since our human neurons are so slow (at least ten million times slower than electronic circuits), we only have time to consider at most a few hundred board positions. Since we are unable to consider the billions of move-countermove situations that a computer such as Deep Blue evaluates, what we do instead is to "deeply" consider each of these situations. So how do we do that? By using *pattern recognition*, which is the heart of human intelligence. We have the ability to recognize situations as being similar to ones we have thought about previously. A chess master such as Kasparov will have mastered up to one hundred thousand such board situations. As he plays, he recognizes situations as being similar to ones he has thought about before and then calls upon his memory of those previous thoughts (e.g., "this is just like that situation that I got into three years ago against grandmaster so-and-so when I forgot to protect my trailing pawn . . .").

I discuss this in my book in order to introduce a proposal to build game-playing machines in a new and hybrid way which would combine the current strength of machines (i.e., the ability to quickly sift through a vast combinatorial explosion of move-countermove sequences) with the more human-like pattern recognition paradigm which represents at least a current superiority of human thinking. Basically, the idea is to use a large (machine-based) neural net to replace the scoring function. Prior to playing, we train that neural net on millions of examples of real-world chess playing (or whatever other game or problem we are addressing). With regard to chess, we have most of the master games of this century on-line, so we can train this extensive neural net on every master game. And then instead of just using an arbitrary set of rules or procedures at the terminal leaves (i.e., the scoring function), we would use this fully trained neural net to make these evaluations. This would combine the combinatorial approach with a pattern recognition approach (which, as I mentioned above, is my area of technical expertise).

I proposed this to Murray Campbell, head of the IBM Deep Blue team, and he was very interested in the idea, and we were going to pursue it, but then IBM cancelled the Deep Blue project. I may yet return to the idea. Recently I brought up the idea again with Campbell.

Searle completely misconstrues this discussion in my book. It is not at all my view that the simple recursive paradigm of Deep Blue is exemplary of how to build flexible intelligence in a machine. The pattern recognition paradigm of the human brain is that solutions emerge from the chaotic and unpredictable interplay of millions of simultaneous processes. And these pattern recognizers are themselves organized in elaborate and shifting hierarchies. In contrast to today's computers, the human brain is massively parallel, combines digital and analog methods, and represents knowledge as highly distributed patterns encoded in trillions of neurotransmitter strengths.

A failure to understand that computing processes are capable of being—just like the human brain—chaotic, unpredictable, messy, tentative, and emergent is behind much of the criticism of the prospect of intelligent machines that we hear from Searle and other essentially materialist philosophers. Inevitably, Searle comes back to a criticism of "symbolic" computing: that orderly sequential symbolic processes cannot recreate true thinking. I think that's true.

But that's not the only way to build machines, or computers.

So-called computers (and part of the problem is the word "computer" because machines can do more than "compute") are not limited to symbolic processing. Nonbiological entities can also use the emergent self-organizing paradigm, and indeed that will be one great trend over the next couple of decades, a trend well under way. Computers do not have to use only 0 and 1. They don't have to be all digital. The human brain combines analog and digital techniques. For example, California Institute of Technology Professor Carver Mead and others have shown that machines can be built by combining digital and analog methods. Machines can be massively parallel. And machines can use chaotic emergent techniques just as the brain does.

My own background is in pattern recognition, and the primary computing techniques that I have used are not symbol manipulation, but rather self-organizing methods such as neural nets, Markov models, and evolutionary (sometimes called genetic) algorithms.

A machine that could really do what Searle describes in the Chinese Room would not be merely "manipulating symbols" because that approach doesn't work. This is at the heart of the philosophical slight of hand underlying the Chinese Room (but more about the Chinese Room below).

It is not the case that the nature of computing is limited to manipulating symbols. Something is going on in the human brain, and there is nothing that prevents these biological processes from being reverse engineered and replicated in nonbiological entities.

Searle writes that "Kurzweil assures us that Deep Blue was actually thinking." This is one of Searle's many out-of-context quotations. The full quotation from my book addresses diverse ways of viewing the concept of thinking, and introduces my proposal for building Deep Blue in a different, more human way:

> After Kasparov's 1997 defeat, we read a lot about how Deep Blue was just doing massive number crunching, not really "thinking" the way his human rival was doing. 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 time to really think very much during the tournament. Mostly he was just drawing upon his mental database of situations he had thought about long ago. Of course, this depends on one's notion of thinking, as I discussed in chapter three. But if the human approach to chess—*neural network based pattern recognition used to identify situations from a library of previously analyzed situations*—is to be regarded as true thinking, then why not program our machines to work the same way? The third way: And that's my idea that I alluded to above as the third school of thought in evaluating the terminal leaves in a recursive search. . . .

Finally, a comment on Searle's view that the "real competition was not between Kasparov and the machine, but between Kasparov and a team of engineers and programmers." Both Deep Blue and Kasparov obtain input and modification to their knowledge bases and strategies from time to time between games. But both Deep Blue and Kasparov use their internal knowledge bases, strategies, and abilities to play each game without any outside assistance or intervention during the game.

On Searle and his Chinese Rooms

John Searle is probably best known for his Chinese Room Argument, which adherents believe demonstrates that machines (i.e., nonbiological entities) can never truly understand anything of significance (such as Chinese). There are several versions of the Chinese Room, of which I will discuss three.

Chinese Room One: A Person and a Computer in a Room

The first involves a person and a computer in a room. I quote here from Professor Searle's 1992 book:

> I believe the best-known argument against strong AI was my Chinese room argument (Searle 1980a) that showed that a system could instantiate a program so as to give a perfect simulation of some human cognitive capacity, such as the capacity to understand Chinese, even though that system had no understanding of Chinese whatever. Simply imagine that someone who understands no Chinese is locked in a room with a lot of Chinese symbols and a computer program for answering questions in Chinese. The input to the system consists in Chinese symbols in the form of questions; the output of the system consists in Chinese symbols in answer to the questions. We might suppose that the program is so good that the answers to the questions are indistinguishable from those of a native Chinese speaker. But all the same, neither the person inside nor any other part of the system literally understands Chinese; and because the programmed computer has nothing that this system does not have, the programmed computer, qua computer, does not understand Chinese either. Because the program is purely formal or syntactical and because minds have mental or semantic contents, any attempt to produce a mind purely with computer programs leaves out the essential features of the mind.

First of all, it is important to recognize that for this system—the person and the computer—to, as Professor Searle puts it, "give a perfect simulation of some human cognitive capacity, such as the capacity to understand Chinese" and to convincingly answer questions in Chinese, this system is essentially passing a Chinese Turing Test. It is entirely equivalent to a Turing Test. In the Turing Test, a computer answers questions in a natural language such as English, or it could be Chinese, in a way that is convincing to a human judge. That is essentially the premise here in the Chinese Room. Keep in mind that we are not talking about answering questions from a fixed list of stock questions (because that's a trivial task), but answering any unanticipated question or sequence of questions from a knowledgeable human interrogator, just as in Turing's eponymous test.

Now, the human in the Chinese Room has little or no significance. He is just feeding things into the computer and mechanically transmitting the output of the computer. And the computer and the human don't need to be in a room either. *Both the human and the room are irrelevant.* The only thing that is significant is the computer.

Now for the computer to really perform this "perfect simulation of a human cognitive capacity, such as the capacity to understand Chinese," it would have to, indeed, understand Chinese. It has, according to the very premise "the capacity to understand Chinese," so it is then entirely contradictory to say that "the programmed computer . . . does not understand Chinese." The premise here directly contradicts itself.

A computer and computer program *as we know them today* could not successfully perform the described task. So if we are to understand the computer to be like today's computers, then it is not fulfilling the premise. The only way that it could fulfill the premise would be for the computer to have the depth and complexity that a human has. That was Turing's brilliant insight in proposing the Turing Test, that convincingly answering questions in a human language really probes all of human intelligence. We're not talking here about answering a question from a canned set of questions, but answering any possible sequence of questions from an intelligent human questioner. A system that could only answer a fixed set of questions would quickly be unmasked by a knowledgeable interlocutor. That requires a human level of intelligence.

A computer that is capable of accomplishing this—a computer that we will run into a few decades from now—will need to be of human complexity or greater, and will indeed understand Chinese in a deep way—because otherwise it would never be convincing in its claim to understand Chinese.

So just stating the computer "does not literally understand Chinese" does not make sense. It contradicts the entire premise.

To claim that the computer is not conscious is not compelling either. To be consistent with some of Searle's other statements, we have to conclude that we really don't know if it is conscious or not. With regard to relatively simple machines, including today's computers, while we can't state for certain that these entities are not conscious, their behavior, including their inner workings, don't give us that impression. But that will not be true for a computer that can really do what is needed in the Chinese room. Such a computer will at least *seem* conscious. Whether it is or not, we really cannot make a definitive statement. But just declaring that it is obvious that the computer (or the entire system of the computer, person and room) is not conscious is far from a compelling argument.

In the quote I read above, Professor Searle is saying that "the program is purely formal or syntactical." But as I pointed out above, that is a bad assumption based on Searle's failure to understand the requirements of such a technology. This assumption is behind much of the criticism of AI that we have heard from certain AI critics such as Searle. A program that is purely formal or syntactical will not be able to understand Chinese, and it won't "give a perfect simulation of some human cognitive capacity."

But again, we don't have to build our machines that way. We can build them the same way nature built the human brain: using chaotic emergent methods that are massively parallel. Furthermore, there is nothing preventing machines from mastering semantics. There is nothing inherent in the concept of a machine that restricts its expertise to the level of syntax alone. Indeed if the machine inherent in Searle's conception of the Chinese Room had not mastered semantics, it would not be able to convincingly answer questions in Chinese and thus would contradict Searle's own premise.

One approach, as I discuss at length in my book and in my chapter in this book, is to reverse engineer and copy the methods of the human brain (with possible extensions). And if it is a Chinese human brain, the copy will understand Chinese. I am not talking about a simulation per se, but rather a duplication of the causal powers of the massive neuron cluster that constitutes the brain, at least those causal powers salient and relevant to thinking.

Will such a copy be conscious? I don't think the Chinese Room Argument tells us anything about this question.

Searle's Chinese Room Argument Can Be Applied to the Human Brain Itself

Although it is clearly not his intent, Searle's own argument implies that the human brain has no understanding. He writes:

> "The computer . . . 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."

Searle acknowledges that biological neurons are machines, so if we simply substitute the phrase "human brain" for "computer" and "neurotransmitter concentrations and related mechanisms" for "formal symbols," we get:

> *The [human brain] . . . succeeds by manipulating [neurotransmitter concentrations and related mechanisms]. The [neurotransmitter concentrations and related mechanisms] themselves are quite meaningless: they have only the meaning we have attached to them. The [human brain] knows nothing of this, it just shuffles the [neurotransmitter concentrations and related mechanisms].*

Of course, neurotransmitter concentrations and other neural details (e.g., interneuronal connection patterns) have no meaning in and of themselves. The meaning and understanding that emerges in the human brain is exactly that: an *emergent* property of its complex patterns of activity. The same is true for machines. Although the "shuffling symbols" do not have meaning in and of themselves, the emergent patterns have the same potential role in nonbiological systems as they do in biological systems such as the brain. As Hans Moravec has written, "Searle is looking for understanding in the wrong places . . . [he] seemingly cannot accept that real meaning can exist in mere patterns."

Chinese Room Two: People Manipulating Slips of Paper

Okay, now let's address a second conception of the Chinese Room. In this conception of the Chinese Room Argument, the room does not include a computer but has a room full of people manipulating slips of paper with Chinese symbols on it. The idea is that this system of a room, people, and slips of paper would convincingly answer questions in Chinese, but none of the participants would know Chinese, nor could we say that the whole system really knows Chinese. Not in a conscious way, anyway. Searle then essentially ridicules the idea that this "system" could be conscious. What are we to consider conscious, Searle asks: the slips of paper, the room? Of course the very notion sounds absurd, so the point is made.

One of the problems with this version of the Chinese Room Argument is that this model of the Chinese Room does not come remotely close to really solving the specific problem of answering questions in Chinese. This form of Chinese Room is really a description of a machine-like process that uses the equivalent of a table look-up, with perhaps some straightforward logical manipulations, to answer questions. It would be able to answer some limited number of canned questions. But if it were to answer *any* arbitrary question that it might be asked, this process would really have to understand Chinese in the same way that a Chinese person does. Again, it is essentially being asked to pass a Chinese Turing Test. And as such, it would need to be as clever, and about as complex, as a human brain, a Chinese human brain. And straightforward table look-up algorithms are simply not going to work.

If we want to recreate a brain that understands Chinese using people as little cogs in the recreation, we would really need a person for each neural connection, so we would need about a hundred trillion people, which means about ten thousand planet Earths with ten billion persons each. This would require a rather large room. And even if extremely efficient organized, this system would run many thousands of times slower than the Chinese brain it is attempting to recreate (by the way, I say thousands, and not millions or trillions because the human brain is very slow compared to electronic circuits—200 calculations per second versus about one billion for machines today).

So Professor Searle is taking an utterly unrealistic solution, one that does not come close to fulfilling its own premise, and then asks us to perform a mental thought experiment that considers whether or not this unrealistic system is conscious, or knows anything about Chinese. The very word "room" is misleading, as it implies a limited number of people with some manageable number of slips of papers. So people think of this so-called "room" and these slips of papers and the rules of manipulating the slips of paper and then are asked to consider if this "system" is conscious. The apparent absurdity of considering this simple system to be conscious is therefore supposed to show that such a recreation of an intelligent process would not really "know" Chinese.

However, if we were to do it right, so that it would actually work, it would take on the order of a hundred trillion people. Now it's true that none of these hundred trillion people would need to know anything about Chinese, and none of them would necessarily know what is going on in this elaborate system. But that's equally true of the neural connections in a real human brain. None of the hundred trillion connections in my brain knows anything about this Discovery Institute book chapter I am writing, nor do any of them know English, nor any of the other things that I know. None of them are conscious of this chapter, nor of any of the things I am conscious of. Probably none of them are conscious at all. But the entire system of them, that is Ray Kurzweil, is conscious. At least, I'm claiming that I'm conscious.

So if we scale up Searle's Chinese Room to be the rather massive "room" it needs to be, who's to say that the entire system of a hundred trillion people simulating a Chinese Brain that knows Chinese isn't conscious? Certainly, it would be correct to say that such a system knows Chinese. And we can't say that it is not conscious anymore than we can say that about any other process. We can't know the subjective experience of another entity (and in at least some of Searle's writings, he appears to acknowledge this limitation). And this massive hundred trillion person "room" is an entity. And perhaps it is conscious. Searle is just declaring ipso facto that it isn't conscious, and that this conclusion is obvious. It may seem that way when you call it a room, and talk about a limited number of people manipulating a limited number of pieces of paper. But as I said, such a system doesn't remotely work.

A key to the philosophical sleight of hand implicit in the Chinese Room Argument has specifically to do with the complexity and scale of the system. Searle says that whereas he cannot prove that his typewriter or tape recorder are not conscious, he feels it is obvious that they are not. Why is this so obvious? At least one reason is because a typewriter and a tape recorder are relatively simple entities.

But the existence or absence of consciousness is not so obvious in a system that is as complex as the human brain, indeed one that may be a direct copy of the organization and causal powers of a real human brain. If such a "system" acts human and knows Chinese in a human way, is it conscious? Now the answer is no longer so obvious. What Searle is saying in the Chinese Room Argument is that we take a simple "machine"—and the conception of a room of people manipulating slips of paper is indeed a simple machine—and then consider how absurd it is to consider such a simple machine to be conscious.

I would agree that a simple machine appears not to be conscious, and that a room of people manipulating slips of paper does not appear to be conscious. But such a simple machine, whether it be a typewriter, a tape recorder, or a room of people manipulating slips of paper cannot possibly answer questions in Chinese. So the fallacy has everything to do with the scale and complexity of the system. Simple machines do not appear to be conscious (again, this is not a proof, but a reasonable conclusion nonetheless). The possible consciousness of machines that are as complex as the human brain is an entirely different question. Complexity alone does not necessarily give us consciousness, but the Chinese Room tells us nothing about whether or not such a system is conscious. The way Searle describes this Chinese Room makes it sound like a simple system, so it seems reasonable to conclude that it isn't conscious. What he doesn't tell you is that the room needs to be much bigger than the solar system, so this apparently simple system isn't really so simple at all.

Chinese Room Three: A Person with a Rule Book

A third variant of the Chinese Room is that there is only one person manipulating slips of papers according to a "rule book." Searle then asks what we are we to consider conscious: the slips of paper, the rule book, the room? Again, the humorous absurdity of the situation clearly implies that the system is not conscious, and does not really "know" Chinese.

But again, it would be utterly infeasible for this little system to provide "a perfect simulation of some human cognitive capacity, such as the capacity to understand Chinese" unless the rule book were to be as complex as a human brain that understands Chinese. And then it would take absurdly long for the human to follow the trillions of rules.

Okay, how about if the rule book simply listed every possible question, and then provided the answer? This would be even less feasible, as the number of possible questions is in the trillions of trillions.

Also keep in mind that the answer to a question would need to consider all of the dialogue that came before it.

The term “rule book” implies a book of hundreds or maybe thousands of pages of rules, but not many trillions of pages.

So again we have a simple machine—a person and a “rule book”—and the apparent absurdity of such a simple system “knowing” Chinese or being conscious. But what really is absurd is the notion that such a system, even in theory, could really answer questions in Chinese in a convincing way.

The version of the Chinese Room Searle cites in his chapter in this book is closest to this third conception. One just replaces “rule book” with “computer program.” But as I point out above, the man in the room is acting like the central processing unit (CPU) of the computer carrying out the program. One could indeed say that the CPU of a computer, being only a small part of a larger system, does not understand what the entire system understands. One has to look for understanding from the right perspective. The understanding, in this case, is distributed across the entire system, including a vastly complex program, and the billions of little notes that the man would have to keep and organize in order to actually follow the program. That’s where the understanding lies, not in the CPU (i.e., the man in the room) alone. It is a distributed understanding embedded in a vast pattern, a type of understanding that Searle appears not to recognize.

Ray Kurzweil’s Chinese Room: With Decorations from the Ming Dynasty

Okay, so here is my conception of the Chinese Room. Call it Ray Kurzweil’s Chinese Room:

There is a human in a room. The room has decorations from the Ming Dynasty. There is a pedestal on which sits a mechanical typewriter. The typewriter has been modified so that there are Chinese symbols on the keys instead of English letters. And the mechanical linkages have been cleverly altered so that when the human types in a question in Chinese, the typewriter does not type the question, but instead types the answer to the question.

Now the person receives questions in Chinese characters, and dutifully presses the appropriate keys on the typewriter. The typewriter types out not the question, but the appropriate answer. The human then passes the answer outside the room.

So here we have a room with a man in it that appears to know Chinese, yet clearly the human does not know Chinese. And clearly the typewriter does not know Chinese either. It is just an ordinary typewriter with its mechanical linkages modified. So despite the fact that the man in the room can answer questions in Chinese, who or what can we say truly knows Chinese? The decorations?

Now you might have some objections to my Chinese Room.

You might point out that the decorations don't seem to have any significance.

Yes, that's true. Neither does the pedestal. The same can be said for the human, and for the room.

You might also point out that the premise is absurd. *Just changing the mechanical linkages in a mechanical typewriter could not possibly enable it to convincingly answer questions in Chinese* (not to mention the fact that we can't fit all the Kanji symbols on the keys).

Yes, that's a valid objection as well. Now the only difference between my Chinese Room conception, and the several proposed by Professor Searle, is that it is patently obvious in my conception that it couldn't possibly work. It is obvious that my conception is absurd. That is not quite as apparent to many readers or listeners with regard to the Searle Chinese Rooms. However, it is equally the case.

Now, wait a second. We can make my conception work, just as we can make Searle's conceptions work. All you have to do is to make the typewriter linkages as complex as a human brain. And that's theoretically (if not practically) possible. But the phrase "typewriter linkages" does not suggest such vast complexity. The same is true when Searle talks about a person manipulating slips of paper or following a book of rules or a computer program. These are all equally misleading conceptions.

The Chinese Room and Chess

Searle's application of his Chinese Room to chess is equally misleading. He says the man in the room "looks up in a book what he is supposed to do." So again, we have a simple look-up procedure. What sort of book is Searle imagining? If it lists all the chess situations that the man might confront, there wouldn't be enough particles in the Universe to list them all, given the number of possible permutations of chess boards. If, on the other hand, the book contains the program that Deep Blue follows, the man would take thousands of years to make a move, which last time I checked, is not regulation chess.

Searle's primary point is contained in his statement:

> 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.

As I pointed out earlier, Searle is simply assuming his conclusion: the man "is just a computer," so obviously (to Searle) he cannot understand anything. But the entire system which includes the rule book and the man following the rule book does "understand" chess, or else it wouldn't be able to play the game.

It should also be pointed out that playing good chess, even championship chess, is a lot easier than convincingly answering questions in a natural human language such as Chinese. But then, Searle shifts the task from playing chess to being knowledgeable about chess in a human context: knowing something about the history and role of chess, having knowledge about the roles of kings and queens who do not necessarily stand on chess squares, having reasons to want to win the game, being able to articulate such reasons, and so on. A reasonable test of such knowledge and understanding of context would be answering questions about chess and engaging in a convincing dialogue (in the Turing Test sense) about chess using a human language such as English or Chinese. And now we have a task that is very similar to the original Chinese Room task, to which my comments above pertain.

On the Difference Between Simulation and Re-Creation

This discussion of Searle's, which he numbers (1), is so hopelessly confused that it is difficult to know where to begin to unravel his tautological and contradictory reasoning.

Let me start with Searle's stomach analogy. He writes:

> 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.

As I point out in at the beginning of my discussion of Searle's chapter above, Searle confuses simulation with functionally equivalent recreation. We could indeed stuff a pizza into an artificial stomach. It may have a very different design than an ordinary human stomach, but if properly designed, it would digest the pizza as well, or perhaps even better than, a real stomach (in the case of some people's stomachs, that probably wouldn't be so hard to do).

In my chapter and in my book, I discuss the creation of functionally equivalent recreations of individual neurons (which has been done), of substantial clusters of neurons (which has also been done), and, ultimately, of the human brain. I am not talking about conventional neural nets, which involve mathematically simplified neurons, but recreations of the full complexity of the digital-analog behavior and response of human and other mammalian neurons and neuron clusters. And these clusters have been growing rapidly (in accordance with the law of accelerating returns). A few years ago, we could only replicate individual neurons, then we could replicate clusters of tens of neurons, then hundreds, and scientists are now replicating clusters of thousands of neurons. Scaling up to the billions of neurons in the human brain may seem daunting, but so did the human genome scan when first proposed.

I don't assume that a perfect or near-perfect recreation of a human brain would necessarily be conscious. But we can expect that it would exhibit the same subtle, complex behavior and abilities that we associate with humans. Our wonderful ability to connect chessboard kings to historical kings and to reflect on the meaning of chess, and all of our other endearing abilities to put ideas in a panoply of contexts is the result of the complex swirl of millions of interacting processes that take place in the human system. If we recreate (and ultimately, vastly extend) these processes, we will get comparably rich and subtle behavior. Such entities will at least convincingly seem conscious. But I am the first to agree that this does not prove that they are in fact conscious.

Searle writes:

> 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.

Here again, Searle assumes that the methods used by Deep Blue are the only way to build intelligent machines. Searle may assume this, but that is clearly not what my book discusses. There are other methods that do not involve the manipulation of formal symbols in this sense. We have discovered that the behavior and functioning of neurons, while quite complex, are describable in mathematical terms. This should not be surprising, as neurons are constructed of real materials following natural laws. And chips have been created that implement these descriptions, and the chips operate in a very similar manner to biological neurons. We are even putting such chips in human brains to replace disabled portions of those brains, as in the case of neural implants for deafness, Parkinson's Disease, and a growing list of other conditions.

There is nothing in Searle's arguments that argues against our ability to scale up these efforts to capture all of human intelligence, and then extend it in nonbiological mediums. As I pointed out above, these efforts are already scaling up very quickly.

Searle writes:

> 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.

It is remarkable that Searle describes the Church-Turing Thesis as a "mathematical result," but more about that later. Searle here is confusing different results of Church and Turing. Turing and Church independently derived mathematical theorems that show that methods such as a neural net can be carried out, albeit very slowly, on a

Turing Machine, which can be considered as a universal symbol-manipulating machine. They also put forth a conjecture, which has become known as the Church-Turing Thesis, which is not mathematical in nature, but rather relates certain abilities of the human brain (in particular its mathematical abilities) to abilities of a Turing Machine.

We know in practical terms that we can precisely replicate neural functioning in electronic devices. No one has demonstrated any practical limits to our ability to do this. In the book, I discuss our efforts to understand the human brain, and the many different schools of thought pursuing the replication of its abilities.

Searle acknowledges that neural nets can be emulated through computation. Well, that only confirms my thesis. Although many contemporary neural nets involve highly simplified models of neurons, a neural net does not necessarily need to be based on such simplified models of biological neurons. They can be built from models of neurons that are just as complex as biological neurons, or even more complex. And doing so would not change the implications of Turing's and Church's theorems. So we could still replicate these neural nets through forms of computation. And indeed we have been successfully doing exactly this, and such efforts are rapidly increasing in complexity.

As for simulation not being duplication, as I pointed out above, I am specifically talking about functionally equivalent duplication.

Searle writes:

> He [Kurzweil] 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.

As I discussed earlier, Searle frequently changes my statements in critical ways, and in this case has left out the word "convincingly." Of course one can trivially make a computer claim to be conscious. I make the same point repeatedly. Claims to be conscious neither prove nor even suggest its actual presence, nor does an inability to make such a claim demonstrate a lack of consciousness. What I am asserting, specifically, is that we will meet entities in several decades that *convincingly* claim to be conscious.

Searle asserts that I assert that people will "just come to accept without question that machines are conscious." This is a typical distortion of Searle. Many people will accept that machines are conscious precisely because the claims will be convincing. There is a huge difference between idle claims (which are feasible today), and *convincing* claims (which are not yet feasible). It is the difference between the twentieth and twenty-first centuries, and one of the primary points of my book.

Now what does it mean to be convincing? It means that when a nonbiological entity talks about its feelings, its behavior at that moment and subsequently will be fully consistent with what we would expect of a human who professed such feelings. This requires enormously subtle, deep, and complex behavior. Nonbiological entities today do not have this ability. What I am specifically claiming is that twenty-first century nonbiological entities will.

This development will have enormous implications for the relationship between humans and the technology we will have created, and I talk extensively in the book about these implications.

One of those implications is not that such entities are necessarily conscious, even though their claims (to be conscious) will be convincing. We come back to the inability to penetrate the subjective experience of another entity. We accept that other humans are conscious, but even this is a shared assumption. And humans are not of like mind when it comes to the consciousness of non-human entities such as animals. We can argue about the issue, but there is no definitive consciousness-detector that we can use to settle the argument. The issue of the potential consciousness of nonbiological entities will be even more contentious than the arguments we have today about the potential consciousness of non-human entities. My prediction is more a political prediction than a philosophical one.

As I mentioned earlier, Searle writes: "Actual human brains cause consciousness by a series of specific neurobiological processes in the brain."

Searle provides (and has provided) no basis for such a startling view. To illuminate where Searle is coming from, I take the liberty of quoting from a letter Searle sent me (dated December 15, 1998), in which he writes

> . . . it may turn out that rather simple organisms like termites or snails are conscious. . .The essential thing is to recognize that consciousness is a biological process like digestion, lactation, photosynthesis, or mitosis, and you should look for its specific biology as you look for the specific biology of these other processes.

I wrote Searle back:

> Yes, it is true that consciousness emerges from the biological process(es) of the brain and body, but there is at least one difference. If I ask the question, "does a particular entity emit carbon dioxide," I can answer that question through clear objective measurement. If I ask the question, "is this entity conscious," I may be able to provide inferential arguments—possibly strong and convincing ones—but not clear objective measurement.

With regard to the snail, I wrote:

> Now when you say that a snail may be conscious, I think what you are saying is the following: that we may discover a certain neurophysiological basis for consciousness (call it "x") in humans such that when this basis was present humans were conscious, and when it was not present humans were not conscious. So we would presumably have an objectively measurable basis for consciousness. And then if we found that in a snail, we could conclude that it was conscious. But this inferential conclusion is just a strong suggestion, it is not a proof of subjective experience on the snail's part. It may be that humans are conscious because they have "x" as well as some other quality that essentially all humans share, call this "y." The "y" may have to do with a human's level of complexity or something having to do with the way we are organized, or with the quantum properties of our tubules (although this may be part of "x"), or something else entirely. The snail has "x" but doesn't have "y" and so it may not be conscious.
>
> How would one settle such an argument? You obviously can't ask the snail. You can't tell from its fairly simple and more-or-less predictable behavior. Pointing out that it has "x" may be a good argument and many people may be convinced by it. But it's just an argument, it's not a direct measurement of the snail's subjective experience. Once again, objective measurement is incompatible with the very concept of subjective experience.
>
> And indeed we have such arguments today. Not about snails so much, but about higher level animals. It is

> apparent to me that dogs and cats are conscious, and I think you mentioned that you accept this as well. But not all humans accept this. I can imagine scientific ways of strengthening the argument by pointing out many similarities between these animals and humans, but again these are just arguments, not scientific proof.

Searle expects to find some clear biological "cause" of consciousness. And he seems unable to acknowledge that either understanding or consciousness may emerge from an overall pattern of activity. Other philosophers, such as Daniel Dennett, have articulated such "pattern emergent" theories of consciousness. But whether "caused" by a specific biological process or by a pattern of activity, Searle provides no foundation for how we would measure or detect consciousness. Finding a neurological correlate of consciousness in humans does not prove that consciousness is necessarily present in other entities with the same correlate, nor does it prove that the absence of such correlate indicates the absence of consciousness. Such inferential arguments necessarily stop short of direct measurement. In this way, consciousness differs from objectively measurable processes such as lactation and photosynthesis.

Searle writes in his chapter: "It is out of the question, for purely neurobiological reasons, to suppose that the chair or the computer is conscious."

Just what neurobiological reasons is Searle talking about? I agree that chairs don't seem to be conscious, but as for computers of the future that have the same complexity, depth, subtlety, and capabilities as humans, I don't think we can rule out the possibility that they are conscious. Searle just assumes that they are not, and that it is "out of the question" to suppose otherwise. There is really nothing more of a substantive nature to Searle's "arguments" than this tautology.

Now part of the appeal of Searle's stance against the possibility of a computer being conscious is that the computers we know today just don't seem to be conscious. Their behavior is brittle and formulaic, even if they are occasionally unpredictable. But as I pointed out above, computers today are still a million times simpler than the human brain, which is at least one reason they don't share all of the endearing qualities of human thought. But that disparity is rapidly shrinking, and will ultimately reverse itself in several decades. The twenty-first century machines I am talking about in the book will appear and act very differently than the relatively simple computers of today.

Searle may assert that the level of complexity and capacity is irrelevant, that even if nonbiological entities become trillions of times more complex and capable than humans, they inherently just don't have this mysterious neurobiological basis of consciousness. I have no problem with his believing that, but he should present it simply as his belief, and not wrap it in tautological arguments that provide no basis for such a belief.

The Chinese Room Argument is based on the idea that it seems ridiculous that a simple machine can be conscious. He then describes a simple machine successfully carrying out deep, extremely complex tasks such as answering unanticipated questions in Chinese. But simple machines would never accomplish such tasks. However, with regard to the extremely complex machines that could accomplish such difficult and subtle tasks, machines that would necessarily match or exceed the complexity of the human brain, the Chinese Room tells us nothing about their consciousness. It may be that consciousness emerges from certain types of very complex self-organizing processes that take place in the human brain. If so, then recreating the essence of these processes would also result in consciousness. It is certainly a plausible conjecture.

Searle writes:

> 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.

Once again, Searle misrepresents the essence of my argument. As I described in my chapter in this book and in greater detail in my book, I describe this slippery-slope scenario and then provide two strong arguments: one that consciousness is preserved, and a second argument that consciousness is not preserved. I present this specifically to illustrate the contradictions inherent in simplistic explanations of the phenomenon of consciousness. The difficulty of resolving this undeniably important issue, and the paradoxes inherent in our understanding of consciousness, stem, once again, from our inability to penetrate subjective experience with objective measurement. I frequently present the perplexity of the issue of consciousness by showing how reasonable and logical arguments lead us to contradictory conclusions. Searle takes one of these arguments completely out of context and then presents that as my position.

As for "abandoning the main thesis of [my] book, Searle's assertion is absurd. The primary thesis of the book is exactly this: that we will recreate the processes in our brains, and then extend them, and ultimately merge these enhanced processes into our human-machine civilization. I maintain that these recreated nonbiological systems will be highly intelligent, and use this term to refer to the highly flexible skills that we exhibit as humans.

On the Difference Between Intrinsic (Observer Independent) and Observer-Relative Features of the World

With regard to Searle's argument that he numbers (2), I will respond briefly as many of the points have already been covered. First of all, I will point out that from a prevalent interpretation of quantum theory, all features of the world are rendered as observer-relative. But let's consider Searle's distinction as valid for the world as it appears to us.

Searle writes:

> 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.

Searle doesn't define what he means by mental life. But by any reasonable interpretation of the term, I would grant that Searle's observation is reasonable with respect to pocket calculators, cars, and the like. The statement is also reasonable with regard to today's computers. But as for the "computers" that we will meet a few decades from now, Searle's statement just reveals, once again, his bias that computers are inherently incapable of "mental life." It is an assumption that produces an identical conclusion, one of Searle's many tautologies.

If by "mental life," Searle is talking about our human ability to place ideas in a rich array of contexts, to deal with subjects in a fluid and subtle way, to recognize and respond appropriately to human emotions, and all of the other endearing and impressive qualities of our species, then computers (nonbiological entities) will achieve—according to the primary thesis of my book—these abilities and behaviors. If we're talking about consciousness, then we run into the same objective-subjective barrier.

Searle writes:

> 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 comes 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 compu-

> tation to commercial machinery, in that it requires an outside observer or interpreter.

There are many confusions in the lengthy quote above, several of which I have already discussed. When I speak of the intelligence that will emerge in twenty-first century machines as a result of reverse engineering the human brain and recreating and extending these extensive processes in extremely powerful new substrates, I am not talking about trivial forms of "intelligence" such as found in calculators and contemporary chess machines. I am not referring to the "narrow" victories of contemporary computers in areas such as chess, diagnosing blood cell images, or tracking land terrain images in a cruise missile. What I am talking about is recreating the processes that take place in the human brain, which, as Searle acknowledges, is a machine that follows natural laws in the physical world. It is disingenuous for Searle to maintain that I confuse the narrow calculations of a calculator or even a game-playing algorithm with the sorts of deep intelligence displayed by the human brain.

I do not maintain that the processes that take place in human brains can be recreated in nonbiological machines because human beings are capable of performing arithmetic. This is typical of Searle's disingenuous arguments: attributing absurd assertions to my book that in fact it never makes, and then pointing to their absurdity.

Another example is his false statement that I assume that the main thing humans do in their lives is compute. I make the opposite point: very little of our time is spent "computing." I make it clear that what goes on in the human brain is a pattern recognition paradigm: the complex, chaotic, and unpredictable interplay of millions of intersecting and interacting processes. We have in fact no direct means of performing mental computation (in the sense that Searle refers to in the above quote) at all. When we perform "computations" such as figuring out 2+2, we use very indirect and complex means. There is no direct calculator in our brains.

A nonbiological entity that contains an extended copy of the very extensive processes that take place in the human brain can combine the resulting human-like abilities with the speed, accuracy and sharing ability that constitute a current superiority of machines. As I mentioned above, humans are unable to directly transfer their knowledge to other persons. Computers, however, can share their knowledge very quickly. As we replicate the functionality of human neuron clusters, we are not leaving out quick downloading ports on the neurotransmitter strength patterns. Thus future machines will be able to combine human intellectual and creative strengths with machine strengths. When one machine learns a skill or gains an insight, it will be able to share that knowledge instantly with billions of other machines.

On the Church-Turing Thesis

Searle makes some strange statements about the Church-Turing Thesis, an important philosophical thesis independently presented by Alan Turing and Alonzo Church.

Searle writes:

> 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.

Searle also writes:

> [T]he basic idea [of the Church-Turing Thesis] 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.

It is remarkable that Searle refers to the Church-Turing Thesis as a "mathematical result." He must be confusing the Church-Turing Thesis (CTT) with Church and Turing theorems. CTT is not a mathematical theorem at all, but rather a philosophical conjecture which relates to a proposed relationship between what a human brain can do and what a Turing Machine can do. There are a range of versions or interpretations of CTT. A standard version is that any method that a human can use to solve a mathematical problem in a finite amount of time can be expressed as a general recursive function and can therefore be solved in a finite amount of time on a Turing Machine. Searle's definition only makes sense if we interpret his phrase "algorithmic solution" to mean a method that a human follows, but that is not the common meaning of this phrase (unless we qualify the phrase as in "algorithmic solutions implemented by a human brain"). The phrase "algorithmic solution" usually refers to a method that can be implemented on a Turing Machine. This makes the Searle definition a tautology.

Broader versions of CTT consider problems beyond mathematical problems, which is consistent with the definition I offer in the book's timeline. The definition I provided is necessarily simplified as it is one brief entry in a lengthy timeline ("1937: Alonzo Church and Alan Turing independently develop the Church-Turing Thesis. This thesis states 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"). In this conception of CTT, I relate problems solvable by a human to algorithms, and use the word "algorithms" in its normal sense as referring to methods that can be implemented on a Turing Machine.

There are yet broader conceptions of CTT that relate the processes that take place in the human brain to methods that are computable. This conjecture is based on the following: (i) the constituent components of brains (e.g., neurons, interneuronal connections, synaptic clefts, neurotransmitter concentrations) are made up of matter and energy, therefore: (ii) these constituent components follow physical laws, therefore: (iii) the behavior of these components are describable in mathematical terms (even if including some irreducibly random elements), therefore: (iv) the behavior of such components is machine-computable.

In Conclusion

I believe that the scale of Searle's misrepresentation of ideas from the AI community stems from a basic lack of understanding of technology. He is stuck in a mindset that nonbiological entities are only capable of manipulating logical symbols, and appears to be unaware of other paradigms. It is true that manipulating symbols is largely how rule-based expert systems and game-playing programs such as Deep Blue work. But the current trend is in a different direction, towards self-organizing chaotic systems that employ biological-inspired methods, including processes derived directly from the reverse engineering of the hundreds of neuron clusters we call the human brain. Searle acknowledges that biological neurons are machines, indeed that the entire brain is a machine. Recent advances that I discussed above have shown that we can recreate in an extremely detailed way the "causal powers" of individual neurons as well as those of substantial neuron clusters. There is no conceptual barrier to scaling these efforts up to the entire human brain.

Searle argues that the Church-Turing Thesis (it's actually Church and Turing theorems) show that neural nets can be mapped onto algorithms that can be implemented in machines. Searle's own argument, however, can be applied equally well to *biological* neural nets, and indeed the experiments I cite above demonstrate this empirically.

Searle is a master of combining tautologies and contradictions in the same argument, but his illogical reasoning to the effect that machines that demonstrate understanding have no understanding does nothing to alter these rapidly accelerating developments.