Gelernter, Kurzweil debate machine consciousness

December 6, 2006

Are we limited to building super-intelligent robotic “zombies” or will it be possible and desirable for us to build conscious, creative, volitional, perhaps even “spiritual” machines? David Gelernter and Ray Kurzweil debated this key question at MIT on Nov. 30.

Transcript by MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), published with permission on KurzweilAI.net December 6, 2006. Participants: Yale professor of computer science David Gelernter, Ray Kurzweil, and CSAIL Director Rodney Brooks, acting as moderator, with questions from the audience.

BROOKS: This is a double-headed event today. We’re going to start off with a debate. Then we’re going—maybe it’s a triple-headed event. We’re going to start off with a debate, then we’re going to have a break for pizza and soda—pizza lover here—outside, and then we’re going to come back for a lecture.

The event that this is around is the 70th anniversary of a paper by Alan Turing, "On Computable Numbers," published in 1936, which one can legitimately, I think—I think one can legitimately think of that paper as the foundation of computer science. It included the invention of the Turing—what we now call the Turing Machine. And Turing went on to have lots of contributions to our field, we at the Computer Science and Artificial Intelligence Lab. In 1948, he had a paper titled, "Intelligent Machinery," which I think is really the foundation of artificial intelligence.

So in honor of that 70th anniversary, we have a workshop going on in the next couple days and this even tonight. This event is sponsored by the Templeton Foundation. Charles Harper of the Templeton Foundation is here, and so is Mary Ann Meyers and some other people sponsoring this event. And Charles, I have to ask you one question —A or B? You have to say. You have to choose. This is going to choose who goes first, but I’m not telling you who A or B is.

HARPER: A.

BROOKS: OK. So we’re going to start this debate between Ray Kurzweil and David Gelernter. And it turns out that Ray is going to go first. Thanks, Charles. So I’m first going to introduce Ray and David. I will point out that after we finish and after the break, we’re going to come

back at 6:15, and Jack Copeland, who's down here, will then give a lecture on Turing's life. And Jack has been—runs the Alanturing.net, the archives in New Zealand of Alan Turing, and he's got a wealth of material and new material that's being declassified over time that he'll be talking about some of Alan Turing's contributions.

But the debate that we're about to have is really about the AI side of Alan Turing and the limits that we can expect or that we might be afraid of or might be celebrating of whether we can build superintelligent machines, or are we limited to building just superintelligent zombies. We're pretty sure we can build programs with intelligence, but will they just be zombies that don't have the real oomph of us humans? Will it be possible or desirable for us to build conscious, volitional, and perhaps even spiritual machines?

So we're going to have a debate. Ray is going to speak for five minutes and then David is going to speak for five minutes—opening remarks. Then Ray will speak for ten minutes, David for ten minutes —that's a total of 30 minutes, and I'm going to time them. And then we're going to have a 15-minute interplay between the two of them. They get to use as much time as they can get from the other one during that. And then we're going to open up to some questions from the audience. But I do ask that when we have the questions, the questions shouldn't be for you to enter the debate. It would be better if you can come up with some question which you think they can argue about, because that's what we're here to see.

Ray Kurzweil has been a well-known name since his—in artificial intelligence since his appearance on Steve Allen's show in 1965, where he played a piano piece that a computer he had built had composed. Ray has gone on to—

KURZWEIL: I was three years old.

BROOKS: He was three years old, yes. Ray has gone on to build the Kurzweil synthesizers that many musicians use, the Kurzweil reading machines, and many other inventions that have gone out there and are in everyday use. He's got prizes and medals up the wazoo. He won the Lemelson Prize from MIT, he won the National Medal of Technology, presented by President Clinton in 1999. And Ray has written a number of books that have been—come out and been very strong sellers on all sorts of questions about our future and the future of robot kind.

David Gelernter is a professor at Yale University, professor of computer science, but he's sort of a strange professor of computer science, in the sense that he writes essays for *Weekly Standard*, *Time*, *Wall Street Journal*, *Washington Post*, *Los Angeles Times*, and many other sorts of places. And I see a few of my colleagues here, and I'm glad they don't write columns for all those places. His research interests include AI, philosophy of mind, parallel distributed systems, visualization, and information management. And you can read all about them with Google if you want to get more details. Both very distinguished people, and I hope we have some interesting things to hear from them. So we'll start with Ray. And five minutes, Ray.

KURZWEIL: OK. Well, thanks, Rodney. You're very good at getting a turnout. That went quickly. [laughter] So there's a tie-in with my tie, which this was given to me by Intel. It's a photomicrograph of the Pentium, which I think symbolizes the progress we've made since Turing's relay-based computer Ultra that broke the Nazi Enigma code and enabled Britain to

win the Battle of Britain. But we've come a long way since then.

And in terms of this 70th anniversary, the course I enjoyed the most here at MIT, when I was here in the late '60s, was 6.253—I don't remember all the numbers, and numbers are important here —but that was theoretical models of computation, and it was about that paper and about the Turing Machine and what it could compute and computable functions and the busy beaver function, which is non-computable, and what computers can do, and really established computation as a sub-field of mathematics and, arguably, mathematics as a sub-field of computation.

So in terms of the debate topic, I thought it was interesting that there's an assumption in the title that we will build superintelligent machines, we'll build superintelligent machines that are conscious or not conscious. And it brings up the issue of consciousness, and I want to focus on that for a moment, because I think we can define consciousness in two ways. We can define apparent consciousness, which is an entity that appears to be conscious—and I believe, in fact, you have to be apparently conscious to pass the Turing test, which means you really need a command of human emotion. Because if you're just very good at doing mathematical theorems and making stock market investments and so on, you're not going to pass the Turing test. And in fact, we have machines that do a pretty good job with those things. Mastering human emotion and human language is really key to the Turing test, which has held up as our exemplary assessment of whether or not a non-biological intelligence has achieved human levels of intelligence.

And that will require a machine to master human emotion, which in my view is really the cutting edge of human intelligence. That's the most intelligent thing we do. Being funny, expressing a loving sentiment—these are very complex behaviors. And we have characters in video games that can try to do these things, but they're not very convincing. They don't have the complex, subtle cues that we associate with those emotions. They don't really have emotional intelligence. But emotional intelligence is not some sideshow to human intelligence. It's really the cutting edge. And as we build machines that can interact with us better and really master human intelligence, that's going to be the frontier. And in the ten minutes, I'll try to make the case that we will achieve that. I think that's more of a 45-minute argument, but I'll try to summarize my views on that.

I will say that the community, AI community and myself, have gotten closer in our assessments of when that will be feasible. There was a conference on my 1999 book, Spiritual Machines, at Stanford, and there were AI experts. And the consensus then—my feeling then was we would see it in 2029. The consensus in the AI community was, oh, it's going to—it's very complicated, it's going to take hundreds of years, if we can ever do it. I gave a presentation—I think you were there, Rodney, as well, at AI50, on the 50th anniversary of the Dartmouth Conference that gave AI its name in 1956. And we had these instant polling devices, and they asked ten different ways when a machine would pass the Turing test—when will we know enough about the brain, when will we have sophisticated enough software, when will a computer actually pass the Turing test. They got the same answer—it was basically the same question, and they got the same answer. And of course it was a bell curve, but the consensus was 50 years, which, at least if you think logarithmically, as I do, that's not that different from 25 years.

So I haven't changed my position, but the AI community is getting closer to my view. And I'll try to explain why I think that's the case. It's because of the exponential power of growth in information technology, which will affect hardware, but also will affect our understanding of the human brain, which is at least one source of getting the software of intelligence.

The other definition of consciousness is subjectivity. Consciousness is a synonym for subjectivity and really having subjective experience, not just an entity that appears to have subjective experience. And fundamentally—and I'll try to make this point more fully in my ten-minute presentation—that's not a scientific concept. There's no consciousness detector we can imagine creating, that you'd slide an entity in—green light goes on, OK, this one's conscious, no, this one's not conscious—that doesn't have some philosophical assumptions built into it. So John Searle would make sure that it's squirting human neurotransmitters—

BROOKS: Time's up.

KURZWEIL: OK. And Dan Dennett would make sure it's self-reflexive. But we'll return to this.

[applause]

BROOKS: David?

GELERNTER: Let's see. First, I'd like to say thanks for inviting me. My guess is that the position I'm representing—the anti-cognitivist position, broadly speaking—is not the overwhelming favorite at this particular site. But I appreciate your willingness to listen to unpopular opinions, and I'll try to make the most of it by being as unpopular as I can. [Laughter]

First, it seems to me we won't even be able to build superintelligent zombies unless we attack the problem right, and I'm not sure we're doing that. I'm pretty sure we're not. We need to understand, it seems to me, in model thought as a whole the cognitive continuum. Not merely one or a discrete handful of cognitive styles, the mind supports a continuum or spectrum of thought styles reaching from focused analytical thought at one extreme, associated with alertness or wide-awakeness, toward steadily less-focused thought, in which our tendency to free-associate increases. Finally, at the other extreme, that tendency overwhelms everything else and we fall asleep.

So the spectrum reaches from focused analysis to unfocused continuous free association and the edge of sleep. As we move down-spectrum towards free association, naturally our tendency to think analogically increases. As we move down-spectrum, emotion becomes more important. I have to strongly agree with Ray on the importance of emotion. We speak of being coldly logical on the one hand, but dreaming on the other is an emotional experience. Is it possible to simulate the cognitive continuum in software? I don't see why not. But only if we try.

Will we ever be able to build a conscious machine? Maybe, but building one out of software seems to me virtually impossible. First, of course, we have to say what conscious means. For my purpose, consciousness means a subjectivity. And Ray's—and consciousness means the presence of mental states that are strictly private, with no visible functions or consequences.

A conscious entity can call up some thought or memory merely to feel happy, to enjoy the memory, be inspired or soothed or angered by the thought, get a rush of adrenaline from the thought. And the outside world needn't see any evidence of all that this act of thought or remembering is taking place.

Now, the reason I believe consciousness will never be built out of software is that where software is executing, by definition we can separate out, peel off a portable layer that can run in a logically identical way on any computing platform—for example, on a human mind. I know what it's like to be a computer executing software, because I can execute that separable, portable set of instructions just as an electronic digital computer can and with the same logical effect. If you believe that you can build consciousness out of software, you believe that when you execute the right sort of program, a new node of consciousness gets created. But I can imagine executing any program without ever causing a new node of consciousness to leap into being. Here I am evaluating expressions, loops, and conditionals. I can see this kind of activity producing powerful unconscious intelligence, but I can't see it creating a new node of consciousness. I don't even see where that new node would be—floating in the air someplace, I guess.

And of course, there's no logical difference between my executing the program and the computer's doing it. Notice that this is not true of the brain. I do not know what it's like to be a brain whose neurons are firing, because there is no separable, portable layer that I can slip into when we're dealing with the brain. The mind cannot be ported to any other platform or even to another instance of the same platform. I know what it's like to be an active computer in a certain abstract sense. I don't know what it's like to be an active brain, and I can't make those same statements about the brain's creating or not creating a new node of consciousness.

Sometimes people describe spirituality—to move finally to the last topic—as a feeling of oneness with the universe or a universal flow through the mind, a particular mode of thought and style of thought. In principle, you could get a computer to do that. But people who strike me as spiritual describe spirituality as a physical need or want. My soul thirsteth for God, for the living God, as the Book of Psalm says. Can we build a robot with a physical need for a non-physical thing? Maybe, but don't count on it. And forget software.

Is it desirable to build intelligent, conscious computers, finally? I think it's desirable to learn as much as we can about every part of the human being, but assembling a complete conscious artificial human is a different project. We might easily reach a state someday where we prefer the company of a robot from Wal-Mart to our next door neighbors or roommates or whatever, but it's sad that in a world where we tend to view such a large proportion of our fellow human beings as useless, we're so hot to build new ones. [laughter]

In a Western world that no longer cares to have children at the replacement rate, we can't wait to make artificial humans. Believe it or not, if we want more complete, fully functional people, we could have them right now, all natural ones. Consult me afterwards, and I'll let you know how it's done. [laughter]

BROOKS: OK, great.

GELERNTER: Thank you.

KURZWEIL: You heard glimpses in David's presentation of both of these concepts of consciousness, and we can debate them both. I think principally he was talking about a form of performance that incorporates emotional intelligence. Because emotional intelligence, even though it seems private and we assume that there is someone actually home there experiencing the emotions that are apparently the case, we can't really tell that when we look at someone else. In fact, all that we can discuss scientifically is objective observation, and science is really a synonym for objectivity, and consciousness is a synonym for subjectivity, and there is an inherent gulf between them.

So some people feel that actual consciousness doesn't exist, since it's not a scientific concept, it's just an illusion, and we shouldn't waste time talking about it. That's not fully satisfactory, in my view, because our whole moral and ethical and legal system is based on consciousness. If you cause suffering to some other conscious entity, that's the basis of our legal code and ethical values. Some people describe some magical or mystical property to consciousness. There were some elements in David's remarks, say, in terms of talking about a new node of consciousness and how that would suddenly emerge from software.

My view is it's an emergent property of a complex system. It's not dependent on substrate. But that is not a scientific view, because there's really no way to talk about or to measure the subjective experience of another entity. We assume that each other are conscious. It's a share human assumption. But that assumption breaks down when we go out of shared human experience. The whole debate about animal rights has to do with are these entities actually conscious. Some people feel that animals are just machines in the old-fashioned sense of that term, not—there's nobody really home. Some people feel that animals are conscious. I feel that my cat's conscious. Other people don't agree. They probably haven't met my cat, but —(laughter)

But then the other view is apparent consciousness, an entity that appears to be conscious, and that will require emotional intelligence. There are several reasons why I feel that we will achieve that in a machine, and that has to do with the acceleration of information technology —and this is something I've studied for several decades. And information technology, not just computation, but in all fields, is basically doubling every year in price, performance, capacity, and bandwidth. We certainly can see that in computation, but we can also see that in other areas, like the resolution of brain-scanning in 3D volume is doubling every year, the amount of data gathering on the brain is doubling every year. And we're showing that we can actually turn this data into working models and simulations of brain regions. There's about 20 regions of the brain that have already been modeled and simulated.

And I've actually had a debate with Tomaso Poggio as to whether this is useful, because he kept saying, well, OK, we'll learn how the visual cortex works, but that's really not going to be useful in creating artificial vision systems. And I said, well, when we got these early transformations of the auditory cortex, that actually did help us in speech recognition. It was not intuitive, we didn't expect it, but when we plugged it into the front-end transformations of speech recognition, we got a big jump in performance. They haven't done that yet in visual modeling of the visual cortex. And I saw him recently—in fact, at AI50—and he said, you know, you were right about that, because now they're actually getting models, these early

models of how the visual cortex works, and that that has been helpful in artificial vision systems.

I make the case in chapter four of my book that we will have models and simulations of all several hundred regions of the human brain within 20 years. And you have to keep in mind that the progress is exponential. So it's very seductive. It looks like nothing is happening. People dismissed the genome project. Now we think it's a mainstream project, but halfway through the project, only 1% of the project had been done, but the amount of genetic data doubled smoothly every year and the project was done on time. If you can factor in this exponential pace of progress, I believe we will have models and simulations of these different brain regions—IBM is already modeling a significant slice of the cerebral cortex. And that will give us the templates of intelligence, it will expand the AI toolkit, and it'll also give us new insights into ourselves. And we'll be able to create machines that have more facile emotional intelligence and that really do have the subtle cues of emotional intelligence, and that will be necessary to passing the Turing test.

But that still doesn't—that still begets the key question as to whether or not those entities just appear to be conscious and feeling emotion or whether they really have emotional subjective experiences. David, I think, was giving a sophisticated version of John Searle's Chinese room argument, where—I don't have time to explain the whole argument, but for those of you familiar with it, you've got a guy that's just following some rules on a piece of paper and he's answering questions in Chinese, and John says, well, isn't it ridiculous to think that that system is actually conscious? Or he has a mechanical typewriter which types out answers in Chinese, but it's following complex rules. The premise seems absurd that that system could actually be—have true understanding and be conscious when it's just following a simple set of rules on a piece of paper.

Of course, the sleight of hand in that argument is that these set of rules would be immensely complex, and the whole premise is unrealistic that such a simple system could, in fact, realistically answer unanticipated questions in Chinese or any language. Because basically what the man is doing in the Chinese room, in John Searle's argument, is passing a Turing test. And that entity would have to be very complex. And in that complexity is a key emergent property. So David says, well, it seems ridiculous to think that software could be conscious or even—and I'm not sure if he's—which flavor of consciousness he's using there, the true subjectivity or just apparent consciousness, but in either case it seems absurd that a little software program could display that kind of complexity and self-emergent awareness.

But that's because you're thinking of software as you know it today, if in fact you have a massively parallel system, as the brain is, with 100 trillion internal connections, all of which are computing simultaneously, and which in fact we can model those internal connections and neurons quite realistically in some cases today. We're still in the early part of that process. But even John Searle agrees that a neuron is basically a machine and can be modeled and simulated, so why can't we do that with massively parallel system with 100 trillion-fold parallelism? And if that seems ridiculous, that is ridiculous today, but it's not ridiculous with the kind of technology we'll have with 30 more doublings of price, performance, capacity, and bandwidth of information technology, the kind of technology we'll have around 2030.

These massively parallel systems with the complexity of the human brain, which is a moderate level of complexity, because the design of the human brain is in the genome and the genome has 800 million bytes, but that's uncompressed, has massive redundancies—ALU's repeated 300,000 times. If you apply loss that's compression of the genome, you can reduce it to 30-50 million bytes, which is not simple, but it's a level of complexity we can manage.

BROOKS: Ray, the logarithm of your remaining time is one. [laughter]

KURZWEIL: So the—we'll be able to achieve that level of complexity. We are making exponential progress in reverse engineering the brain. We'll have systems that have the suppleness of human intelligence. This will not be conventional software as we understand it today. There is a difference in the (inaudible) field of technology when it achieves that level of parallelism and that level of complexity, and I think we'll achieve that if you consider these exponential progressions. And it still doesn't penetrate the ultimate mystery of how consciousness can emerge, true subjectivity. We assume that each other are conscious, but that assumption breaks down in the case of animals, and we'll have a vigorous debate when we have these machines. But I'll make one point. We will—I'll make a prediction that we will come to believe these machines, because they'll be very clever and they'll get mad at us if we don't believe them, and we won't want that to happen. So thank you.

BROOKS: OK. David?

GELERNTER: Well, thank you for those very eloquent remarks. And I want to say, first of all, many points were raised. The premise of John Searle's Chinese room and of the thought experiment which is related, that I outlined, is certainly unrealistic. Granted, the premise is unrealistic. That's why we have thought experiments. If the premise were not unrealistic, if it were easy to run in a lab, we wouldn't need to have a thought experiment.

Now, the fact remains that when we conduct a thought experiment, any thought experiment needs to be evaluated carefully. The fact that we can imagine something doesn't mean that what we imagine is the case. We need to know whether our thought experiment is based on experience. I would say the thought experiment of imagining that you're executing the instructions that constitute a program or that realize a virtual machine is founded on experience, because we've all had the experience of executing algorithms by hand. It isn't any—and there's no exotic ingredient in executing instructions. I may be wrong. I don't know for sure what would happen if I executed a truly enormous program that went on for billions of pages. But I don't have any reason for believing that consciousness would emerge. It seems to me a completely arbitrary claim. It might be true. Anything might be true. But I don't see why you make the claim. I don't see what makes it plausible.

You mentioned massive parallelism, but massive parallelism, after all, adds absolutely zero in terms of expressivity. You could have a billion processors going, or ten billion or ten trillion or 1081, and all those processors could be simulated on a single jalopy PC. I could run all those processes asynchronously on one processor, as you know, and what I get from parallelism is performance, obviously, and a certain amount of cleanliness and modularity when I write the program, but I certainly don't get anything in terms of expressivity that I didn't have anyway.

You mentioned consciousness, which is the key issue here. And you pointed out consciousness is subjective. I'm only aware of mine, you're only aware of yours, granted. You say that consciousness is an emergent property of a complex system. Granted, of course, the brain is obviously a complex system and consciousness is clearly an emergent property. Nobody would claim that one neuron tweezed out of the brain was conscious. So yes, it is an emergent property. The business about animals and people denying animal consciousness, I haven't really heard that since the 18th century, but who knows, maybe there are still Cartesians out there—raise your hands.

But in the final analysis, although it's true that consciousness is irreducibly subjective, you can't possibly claim to understand the human mind if you don't understand consciousness. It's true that I can't see yours and you can't see mine. It doesn't change the fact that I know I'm conscious and you know that you are. And I'm not going to believe that you understand the human mind unless you can explain to me what consciousness is, how it's created and how it got there. Now, that doesn't mean that you can't do a lot of useful things without being —creating consciousness. You certainly can. If your ultimate goal is utilitarian, forget about consciousness. But if your goals are philosophical and scientific and you want to understand how the mind really operates, then you must be able to tell me how consciousness works, or you don't have a theory of the human mind.

One element that I think you left out in your discussion of the thought experiment and the fact that, granted, we're able to build more and more complex systems and they are more and more powerful, and we're able to build more and more accurate and effective simulations of parts of the brain and indeed of other parts of the body—because keep in mind that when we allow the importance of emotion and thinking, it's clear that you don't just think with your brain, you think with your body. When you feel an emotion, when you have an emotion, the body acts as a resonator or a sounding board or an amplifier, and you need to understand how the body works, as well as the brain does, if you're going to understand emotion. But granted, we're getting—we're able to build more complex and more and more effective simulators.

What isn't clear is the role of the brain's chemical structure. The role of the brain stuff itself, of course, is a point that Searle harps on, but it goes back to a paper by Paul Ziff in the late 1950s, and many people have remarked on this point. We don't have the right to dismiss out of hand the role of the actual chemical makeup of the brain in creating the emergent property of consciousness. We don't know whether it can be created using any other substance. Maybe it can't and maybe it can. It's an empirical question.

One is reminded of the famous search that went on for so many centuries for a substitute source of the pigment ultramarine. Ultramarine, a tremendously important pigment for any painter. You get it from lapis lazuli, and there are not very many sources of lapis lazuli. It's very expensive, and it's a big production number to get it and grind it down, turn it into ultramarine. So ultramarine paint used to be as expensive as gold leaf. People wanted to know, where else can I get ultramarine? And they went to the scientific community, and the scientific community said, we don't know. There's no law that says there is some other place to get ultramarine from lapis lazuli, but we'll try. And at a certain point in the late 19th century, a team of French chemists did succeed in producing a fake ultramarine pigment which was indeed much cheaper than lapis lazuli. And the art world rejoiced.

The moral of the story? If you can do it, great, but you have no basis for insisting on an a priori assumption that you can do it. I don't know whether there is a way to achieve consciousness in any way other than living organisms achieve it. If you think there is, you've got to show me. I have no reason for accepting that a priori. And I think I'm finished.

BROOKS: I can't believe it. Everyone stopped—Ray, I think—stay up there, and we'll—now we'll go back and forth in terms of, Ray, maybe you want to answer that.

KURZWEIL: So I'm struggling as I listen to your remarks, David, to really tell what you mean by consciousness. I've tried to distinguish these two different ways of looking at it—the objective view, which is usually what people lapse into when they talk about consciousness. They talk about some neurological property, or they talk about self-reflection, an entity that can create models of its own intelligence and behavior and model itself, or what-if experiments in its mind or have imagination, thinking about itself and transforming models of itself and this kind of self-reflection. That is consciousness. Or maybe it has to do with mirror neurons and that we can empathize—that is to say, understand the conscious or the emotions of somebody else.

But that's all objective performance. And these—our emotional intelligence, our ability to be funny or be sad or express a loving sentiment, those are things that the brain does. And I'd make the case that we are making progress, exponential progress in understanding the human brain and different regions, and modeling them in mathematical terms and then simulating them and testing those simulations. And the precision of those simulations is gearing up. We can argue about the timeframe. I think, though, within a quarter century or so, we will have detailed models that—and simulations that can then do the same things that the brain does apparently. And we won't be able to really tell them apart.

That is what the Turing test is all about, that this machine will pass the Turing test. But that is an objective test. We could argue about the rules. Mitch Kapor and I argued for three months about the rules. Turing wasn't very specific about them. But it is a objective test and it's an objective property. So I'm not sure if you're talking about that or talking about the actual sense one has of feeling, your apparent feelings, the subjective sense of consciousness. And so you talk about—

GELERNTER: (inaudible), could I answer that question?

BROOKS: Yeah, let (inaudible).

GELERNTER: You say there are two kinds of consciousness, and I think you're right. I think most people, when they talk about consciousness, think of something that's objectively visible. As I said, for my purposes, I want consciousness to mean mental states, mental states —specifically a mental state that has no external functionality.

KURZWEIL: But that's still—

GELERNTER: You know that you are capable of feeling or being happy. You know you're capable of thinking of something good that makes you feel good, of thinking of something bad that makes you depressed, or thinking of something outrageous that makes you angry.

You know you're capable of mental states that are your property alone. As you say, there's objective—absolutely—

KURZWEIL: But these mental states do have—

GELERNTER: That's what I mean by consciousness.

KURZWEIL: But these mental states still have objective neurological correlates. And in fact, we now have means of where we can begin to look inside the brain with increasing resolution—strike doubling in 3D volume every year—to actually see what's going on in the brain. So sitting there quietly, thinking happy thoughts and making myself happy, you can—there are actually things going on inside the brain, we're able to see them. And so now this supposedly subjective mental state is, in fact, becoming an objective behavior. Not—

GELERNTER: Can I comment on that? I think you're—I think the idea that you're arguing with Descartes is a straw man approach. I don't think anybody argues anymore that the mind is a result of mind stuff, some intangible substance that has no relation to the brain. By arguing that consciousness is objective—I'm agreeing with you that consciousness is objective—I'm certainly not denying that it's created by physical mechanisms. I'm not claiming there's some magical or transcendental metaphysical property. But that doesn't change the fact that in terms of the way you understand it and perceive it, your experiences of it is subjective. That was your term, and I'm agreeing with you. And that doesn't change the fact that it is created by the brain.

Clearly, we're reaching better and better understandings of the brain and of everything else. You've said that a few times, and I certainly don't disagree. The fact that we're getting better and better doesn't mean that necessarily we're going to reach any arbitrary goal. It depends on our methods. It depends if we understand the problem the right way. It depends if we're taking the right route. It seems to me that consciousness is necessary. Unless we understand consciousness as this objective phenomenon that we're all aware of, our brain simulators haven't really told us anything fundamental about the human mind. Haven't told us what I want to know.

KURZWEIL: I think our brain simulators are going to have to work not just the level of the Turing test, but at the level of measuring the objective neurological correlates of these supposedly internal mental states. And there's some information processing going on when we daydream and we think happy thoughts or sad thoughts or worry about something. There's same kinds of things going on as when we do more visibly intelligent tasks. We're, in fact, more and more able to penetrate that by seeing what's going on and modeling these different regions of the brain, including, say, the spindle cells and the mirror neurons, which are involved with things like empathy and emotion—which are uniquely human, although a few other animals have some of them—and really beginning to model this.

We're at an early stage, and it's easy to ridicule the primitiveness of today's technology, which will also always appear primitive compared to what will be feasible, given the exponential progression. But these internal mental states are, in fact, objective behaviors, because we will need to expand our definition of objective behavior to the kinds of things that we can see when we look inside the brain.

GELERNTER: If I could comment on that? If your tests are telling us that they are unable to distinguish that the same thing creates, on the one hand, a mental state of sharply-focused, in which I'm able to concentrate on a problem without my mind drifting and solving it—there's no way to distinguish that mental state from a mental state in which my mind is wandering, I am unable to focus or concentrate on what I'm doing, and then I start dreaming. In fact, cognitive psychologists have found out that we start dreaming and then we fall asleep. If your tests or your simulators are unable to distinguish between the mental state of dreaming or continuous free association on the one hand and focused logical analytic problem-solving on the other, then I think you're just telling us that your tests have failed, because we know that these states are different and we want to know why they're different. It doesn't do any good to say, well, they're caused in the same way. We need to explain the difference that we can observe.

BROOKS: Can I ask a question which I think gets at what this disagreement is? Then I'll ask you two different questions. The question for David is, what would it take to convince you so that you would accept that you could build a conscious computer built on digital substrate? And Ray, what would it take to convince you that digital stuff isn't good enough, we need some other chemicals or something else that David talked about?

KURZWEIL: To answer it myself, I wouldn't get too hung up on digital, because, in fact, the brain is not digital. The neurotransmitters are kind of a digitally-controlled analog phenomena. But when we figure out the salient—the important thing is to figure out what is salient and how information is modeled and what these different regions are actually doing to transform information.

The actual neurons are very complex. There's lots of things going on, but we find out in the—one region of the auditory cortex is basically conducting a certain type of algorithm, the information is represented perhaps by the location of certain neurotransmitters in relation to another, whereas in another case it has to do with the production of some unique neurotransmitter. There's different ways in which the information is represented. And these are chemical processes, but we can model really anything like that at whatever level of specificity is needed digitally. We know that. We can model it analog—

BROOKS: OK, so you didn't answer the question. Can you then answer the question? (laughter)

GELERNTER: I will continue in exactly the same spirit, by not answering the question. I wish I could answer the question. It is a very good question and a deep question. Given the fact that mental states that are purely private are also purely subjective, how can we tell when they are present? And the fact is, just as you don't know how to produce them, I don't know how to tell whether they are there. It's a research question, it's a philosophical question.

It's—we know how to understand particular technologies. That is, we say I've created consciousness and I've done it by running software on a digital computer. I can think about that and say I don't buy that, I don't believe there's consciousness there. If you wheel in some other technology, my only stratagem is to try and understand that new technology. I need to understand what you're doing, I need to understand what moves you're making, because unfortunately I don't know of any general test. The only test that one reads about or hears

about philosophically is relevant similarity—that is, we assume that our fellow human beings are conscious, because we can see they're people like us. We assume that if I had mental states, other similar creatures have mental states. And we make that same assumption about animals. And the more similar to us they seem, the more we assume their mental states are like ours.

How are we going to handle creatures who are—or things or entities, objects, that are radically unlike us and are not organic? It's a hard question and an interesting question. I'd like to see more work done on it.

KURZWEIL: In some ways, they'll be more like us than animals, because animals are not perfect models of humans either medically or mentally. Whereas as we really reverse-engineer what's going on, the salient processes, and learn what's important in the different regions of the brain and recreate those properties and abilities to transform information similar ways, and then get an entity that in fact acts very human-like and a lot more human-like than an animal, for example, can pass a Turing test, which involves mastery of language which animals basically don't have, for the most part, they will be closer to humans than animals are.

If we really model—take an extreme case. I don't think this is necessary to model neuron by neuron and neurotransmitter by neurotransmitter, but one could in theory do that. And we have, in fact, do have simulations of neurons that are highly detailed already, of one neuron or a cluster of three or four of them. So why not extend that to 100 billion neurons? It's theoretically possible, and it's a different substrate, but it's really doing the same things. And it's closer to humans than animals are.

BROOKS: So while David responds, if people who want to ask questions can come to the two microphones. Go ahead.

GELERNTER: When you say act very human-like, this is a key issue. You have to keep in mind that the Turing test is rejected by many people, and has been from the very beginning, as a superficial test of performance, a test that fails to tell us anything about mental states, fails to tell us the things that we really most want to know. So when you say something acts very human-like, that's exactly what we don't do when we attribute the presence of consciousness on the basis of relevant similarity.

When I see somebody, even if he isn't acting human-like at all, if he's fast asleep, even if he's out cold, I don't need to see him do anything, I don't need to have him answer any fancy questions on the Turing test. I can see he's a creature like I am, and I therefore attribute to him a mind and believe he's capable of mental states. On the other hand, the Turing test, which is a test of performance rather than states of being, has been—has certainly failed to convince people who are interested in what you would call the subjective kind of consciousness.

KURZWEIL: Well, I think now we're—

GELERNTER: That doesn't tell me anything about—

KURZWEIL: Well, now I think we're getting somewhere, because I would agree. The Turing test is an objective test. And we can argue about making it super-rigorous and so forth, but—and if an entity passed that test, the super-rigorous one, it is really convincingly human. It's convincingly funny and sad, and we really—is really displaying those emotions in a way that we cannot distinguish from human beings. But you're right—I mean, this gets back to a point I made initially. That doesn't prove that that entity is conscious, and we don't absolutely know that people are conscious. I think we will come to accept them as conscious. That's a prediction I can make. But fundamentally, this is the underlying ontological question.

There is actually a role for philosophy, because it's not fundamentally a scientific question. If you reject the Turing test or any variant of it, then we're just left with this philosophical issue. My own philosophical take is if an entity seems to be conscious, I would accept its consciousness. But that's a philosophical and not a scientific position.

BROOKS: So I think we'll take the first question. And remember, not a monologue, something to provoke discussion.

M: Yeah, no problem. Let's see. What if everything is conscious and connected, and it's just a matter of us learning how to communicate and connect with it?

KURZWEIL: That's a good point, because we can communicate with other humans, to some extent—although history is full of examples where we dehumanize a certain portion of the population and don't really accept their conscious experience—and we have trouble communicating with animals, so that really underlies the whole animal rights— what's it like to be a giant squid? Their behavior seems very intelligent, but it's also very alien and we don't—there's no way we can even have the terminology to express that, because it's not experiences that are human. And that is part of the deep mystery of consciousness and gets at the subjective aspects of it.

But as we do really begin to model our own brain and then extend that to other species, as we're doing with the genome—we're now trying to reverse-engineer the genome in other species, and we'll do the same thing ultimately with the brain—that will give us more insight. We can translate into our own human terms the kinds of mental states as we can see them manifest as we really understand how to model other brains.

GELERNTER: If we think we are communicating with a software-powered robot, we're kidding ourselves, because we're using words in a fundamentally different way. To use an example that Turing himself discusses, we could ask the computer or the robot, do you like strawberries, and the computer could lie and say yes or it could, in a sense, tell the truth and say no. But the more fundamental issue is that not only does it not like strawberries, it doesn't like anything. It's never had the experience of liking, it's never had the experience of eating. It doesn't know what a strawberry is or any other kind of berry or any other kind of fruit or any other kind of food item. It doesn't know what liking is, it doesn't know what hating is. It's using words in a purely syntactic way with no meanings behind them.

KURZWEIL: This is now the Searlean argument, and John Searle's argument can be really rephrased to prove that the human being has no understanding and no consciousness, because each neuron is just a machine. Instead of just shuffling symbols, it's just shuffling

chemicals. And obviously, just shuffling chemicals around is no different than shuffling symbols around. And if shuffling chemicals and symbols around doesn't really lead to real understanding or consciousness, then why isn't that true for a collection of 100 neurons, which are all just little machines, or 100 billion?

GELERNTER: There's a fundamental distinction, which is software. Software is the distinction. I can't download your brain onto the computer up there—

KURZWEIL: Well, that's just a limitation of my brain, because we don't have—we don't have quick downloading ports.

GELERNTER: You need somebody else's brain in the audience?

KURZWEIL: No, that's something that biology left out. We're just not going to leave that out of our non-biological base.

GELERNTER: It turns out to be an important point. It's the fundamental issue—

KURZWEIL: It's a limitation, not—

GELERNTER: I think there's a very big difference whether I can take this computer and upload it to a million other computers or to machines that are nothing like this digital computer, to a Turing machine, to an organic computer, to an optical computer. I can upload it to a class full of freshmen, I can upload it to all sorts of things. But your mind is yours and will never be downloaded (multiple conversations; inaudible)—

KURZWEIL: That's just because we left—

GELERNTER: It's stuck to your brain.

KURZWEIL: We left out the—

GELERNTER: And I think that's a thought-provoking fact. I don't think you can just dismiss it as an—

KURZWEIL: You're posing that as a—

GELERNTER:—envir—a developmental accident. Maybe it is, but—

KURZWEIL: You're posing that as a benefit and advantage of biological intelligence, that we don't have these quick downloading ports to access information—

GELERNTER: Not an advantage. It's just a fact.

KURZWEIL: But that's not an advantage. If we added quick downloading ports, which we will add to our non-biological brain emulations, that's just an added feature. We could leave it out. But we put it in there, that doesn't deprive it of any capability that it would otherwise have.

GELERNTER: You think you could upload your mind to somebody with a different body, with a different environment, who had a different set of experiences, who had a different set of books, feels things in a different way, has a different set of likes, responds in a different kind of way, and get an exact copy of you? I think that's a naïve idea. I don't think there's any way to upload your mind anywhere else and that lets you upload your entire being, including your body.

KURZWEIL: Well, it's hard to upload to another person who already has a brain and a body that's—it's like trying to upload to a machine that's incompatible. But ultimately we will be able to gather enough data on a specific brain and simulate that, including our body and our environmental influences.

BROOKS: Next question.

M: Thanks. If we eventually develop a machine which appears intelligent, and let's say given appropriate body so that it can answer meaningful questions about how does a strawberry taste or something like that or whether it likes strawberries, if we are wondering if this machine is actually experiencing consciousness the same way that we do, why not just ask it? They'll presumably have no reason to lie if you haven't specifically gone out of your way to program that in.

KURZWEIL: Well, that doesn't tell us anything, because we can ask it today. You can ask a character in a video game and it will say, well, I'm really angry or I'm sad or whatever. And we don't believe it, because it doesn't—it's not very convincing yet. It doesn't—because it doesn't have the subtle cues and it's not as complex and not a realistic emulation of—

M: Well, if we built 1000 of them, let's say—

GELERNTER: I strongly agree with (inaudible)—

M:—presumably they wouldn't all agree to lie ahead of time. Somebody—one of them might tell us the truth if the answer is no.

BROOKS: We'll finish that question (multiple conversations; inaudible)—

GELERNTER: I strongly agree. Keep in mind that the whole basis of the Turing test is lying. The computer is instructed to lie and pass itself off as a human being. Turing assumes that everything it says will be a lie. He doesn't talk about the real deep meaning of lying, or he doesn't care about that, and that's fine, that's not his topic. But he'd—it's certainly not the case that the computer is in any sense telling the truth. It's telling you something about its performance, not something about facts or reality or the way it's made or what its mental life is like.

KURZWEIL: John Searle, by the way, thinks that a snail could be conscious if it had this magic property, which we don't understand it, that causes consciousness. And when we figure it out, we may discover that snails have it. That's his view. So I do think that—

GELERNTER: Do you think it's inherently implausible that we should need a certain chemical to produce a certain result? Do you think chemical structure is irrelevant?

KURZWEIL: No, but we can simulate chemical interactions. We just simulated the other day something that people said will never be able to be simulated, which is protein folding. And we can now take an arbitrary amino acid sequence and actually simulate and watch it fold up, and it's an accurate simulation (multiple conversations; inaudible)

GELERNTER: You understand it, but you don't get any amino acids out. As Searle points out, if you want to talk Searlean, you can simulate photosynthesis and no photosynthesis takes place. You can simulate a rainstorm, nobody gets wet. There's an important distinction. Certainly you're going to understand the process, but you're not going to produce the result —

KURZWEIL: Well, if you simulate creativity, you'll—if you simulate creativity, you'll get real ideas out.

BROOKS: Next—sure.

M: So up until this point, there seems to have been a lot of discussion just about a fully—just software, just a human or whatnot. But I'm kind of curious your thoughts towards more of a gray area, if it's possible. That is, if we in some way augment the brain with some sort of electronic component, or somebody has some sort of operation to add something to them. I don't think it's been done yet today, but just is it possible to have fully—what you would consider to be a fully conscious human take part of the brain out, say, replace it with something to do a similar function, and then have obviously the person still survive. Is that person conscious? Is it (inaudible)?

KURZWEIL: Absolutely. And we've done things like that, which I'll mention. But I think—in fact, that is the key application or one key application of this technology. We're not just going to create these superintelligent machines to compete with us from over the horizon. We're going to enhance our own intelligence, which we do now with the machines in our pockets—and when we put them in our bodies and brains, we'll enhance our bodies and brains with them.

But we are applying this for medical problems. You can get a pea-sized computer placed in your brain or placed at biological neurons (inaudible) Parkinson's disease. And in fact, the latest generation now allows you to download new software to your neural implant from outside the patient, and that does replace the function of the corpus of biological neurons. And now you've got biological neurons in the vicinity getting signals from this computer where they used to get signals from the biological neurons, and this hybrid works quite well. And there's about a dozen neural implants, some of which are getting more and more sophisticated, in various stages of development.

So right now we're trying to bring back "normal" function, although normal human function is in fact a wide range. But ultimately we will be sending blood cell-sized robots to the bloodstream non-invasively to interact with our biological neurons. And that sounds very fantastic. I point out there's already four major conferences on blood cell-sized devices that can produce therapeutic functions in animals and—we don't have time to discuss all that, but we will—

BROOKS: Let's hear David's response.

GELERNTER: When you talk about technological interventions that could change the brain, it's a remarkable—it's a fascinating topic, and it can do a lot of good. And one of the really famous instances of that is the frontal lobotomy, an operation invented in the 1950s or maybe the last 1940s. Made people feel a lot better, but somehow it didn't really catch on, because it bent their personality out of shape. So the bottom line is not everything that we do, not every technological intervention that affects your mental state is necessarily going to be good.

Now, it is a great thing to be able to come up with something that cures a disease, makes somebody feel better. We need to do as much of that as we can, and we are. But we—it's impossible to be too careful when you fool around with consciousness. You may make a mistake that you will regret. And lobotomy cases are undoable.

BROOKS: I'm afraid this is going to be the last question.

M: How close do the brain simulation people know they are to the right architecture, and how do they know it? You made the assertion that you don't need to simulate the neurons in detail, and that the IBM people are simulating a slice of neocortex and that's good. And I think that is good, but do they have a theory that says this architecture good, this architecture not good enough? How do they measure it?

KURZWEIL: Well, say, in the case of the simulation of a dozen regions of the auditory cortex done on the West Coast, they've applied sophisticated psychoacoustic tests to the simulation and they get very similar results as applying the same test to human auditory perception. There's a simulation of the cerebellum where they apply skill formation tests. It doesn't prove that these are perfect simulations, but it does show it's on the right track. And the overall performance of these regions appears to be doing the kinds of things that we can measure, that the biological versions do. And the scale and sophistication and resolution of these simulations is scaling up.

The IBM one on the cerebral cortex is actually going to do it neuron by neuron and ultimately at the chemical level, which I don't believe is actually necessary when we—ultimately, to actually create those functions, when we learn the salient algorithms, we can basically implement them using our computer science methods more efficiently. But that's a very useful project to really understand how the brain works.

GELERNTER: I'm all in favor of neural simulations. I think one should keep in mind that we don't think just with our brains, we think with our brains and our bodies. Ultimately, we'll have to simulate both. And we also have to keep in mind that unless our simulators can tell us not only what the input/output behavior of the human mind is, but how it understands and how it produces consciousness —unless it can tell us where consciousness comes from, it's not enough to say it's an emergent phenomenon. Granted, but how? How does it work? Unless those questions are answered, we don't understand the human mind. We're kidding ourselves if we think otherwise.

BROOKS: So with that, I think I'd like to thank both Ray and David. [applause]