From left: W. Daniel Hillis, Neil Gershenfeld, Frank Wilczek, David Chalmers, Robert Axelrod, Tom Griffiths, Caroline Jones, Peter Galison, Alison Gopnik, John Brockman, George Dyson, Freeman Dyson, Seth Lloyd, Rod Brooks, Stephen Wolfram, Ian McEwan. In absentia: Andy Clark, George Church, Daniel Kahneman, Alex "Sandy" Pentland (Click to expand photo)

Edge

Possible Minds Conference

INTRODUCTION
by Venki Ramakrishnan

The field of machine learning and AI is changing at such a rapid pace that we cannot foresee what new technical breakthroughs lie ahead, where the technology will lead us or the ways in which it will completely transform society. So it is appropriate to take a regular look at the landscape to see where we are, what lies ahead, where we should be going and, just as importantly, what we should be avoiding as a society. We want to bring a mix of people with deep expertise in the technology as well as broad

thinkers from a variety of disciplines to make regular critical assessments of the state and future of AI.
—Venki Ramakrishnan, President of the Royal Society and Nobel Laureate in Chemistry, 2009, is Group Leader & Former Deputy Director, MRC Laboratory of Molecular Biology; Author, *Gene Machine: The Race to Decipher the Secrets of the Ribosome.*

[ED. NOTE: In recent months, *Edge* has published the fifteen individual talks and discussions from its two-and-a-half-day Possible Minds Conference held in Morris, CT, an update from the field following on from the publication of the group-authored book *Possible Minds: Twenty-Five Ways of Looking at AI*. As a special event for the long Thanksgiving weekend, we are pleased to publish the complete conference—10 hours plus of audio and video, as well as this downloadable PDF of the 77,500-word manuscript. Enjoy.]

John Brockman
Editor, *Edge*

Table of Contents

IAN MCEWAN
Machines Like Me

I would like to set aside the technological constraints in order to imagine how an embodied artificial consciousness might negotiate the open system of human ethics—not how people think they should behave, but how they do behave. For example, we may think the rule of law is preferable to revenge, but matters get blurred when the cause is just and we love the one who exacts the revenge.

A machine incorporating the best angel of our nature might think otherwise. The ancient dream of a plausible artificial human might be scientifically useless but culturally irresistible. At the very least, the quest so far has taught us just how complex we (and all creatures) are in our simplest actions and modes of being. There's a semi-religious quality to the hope of creating a being less cognitively flawed than we are.

IAN MCEWAN is a novelist whose works have earned him worldwide critical acclaim. He is the recipient of the Man Booker Prize for *Amsterdam* (1998), the National Book Critics' Circle Fiction Award, and the *Los Angeles Times* Prize for Fiction for *Atonement* (2003). His most recent novel is *Machines Like Me*.

* * * *

IAN MCEWAN: I feel something like an imposter here amongst so much technical expertise. I'm the breakfast equivalent of an after-dinner mint.

What's been preoccupying me the last two or three years is what it would be like to live with a fully embodied artificial consciousness, which means leaping over every difficulty that we've heard described this morning by Rod Brooks. The building of such a thing is probably scientifically useless, much like putting a man on the moon when you could put a machine there, but it has an ancient history.

Then of course you had Frankenstein's monster, which shifted the debate into what it means to conjure up a version of ourselves. Now, you have the contemporary TV series of *Westworld* and movies like *Blade Runner* specifically addressing the notion of what it would be like to have an artificial being aware of its own mortality. In medieval churches or cathedrals, you will find wax effigies of the Virgin Mary that, on certain occasions, weep or shed blood. As anyone who's been on the Kurfürstendamm in Berlin will know, there's a Virgin Mary that bleeds. Throughout the 18th century you had water-powered android figures, figures driven by levers and cogs, and as clockwork got more sophisticated in the 18th century, such figures remained a matter of profound interest and fascination.

I've been thinking about what it would be like to live alongside someone we made who is artificial and who claims to have consciousness, about which we'd be very skeptical and to which we'd be applying a constant form of Turing tests. Since it behaved as if it had consciousness, would we then have to accept it much as we have to accept it amongst each other? I've written a novel which takes as a starting point the delivery of such a machine. The year is 1982. Alan Turing, on advice of a close friend, decides that he should not proceed if found guilty on homosexual charges to go for chemical castration and instead does one year in Wandsworth prison.

Cut away from wet bench work, he returns to pure mathematics. He said at that point in his life he was very interested in Dirac. He thought that quantum mechanics had been largely neglected because of the war, and so he sets out to solving, although it's formulated differently, positively P versus NP which, along with various other factors, puts science, robotics, and AI in a different position than it is. Through this novel, I want to reflect on the fragility of the present.

It seems the way things are is the way they were always bound to be, but the éminence grise of this novel is Turing at the age of seventy. He is head of a very large corporation, an outfit rather like Demis Hassabis' Deep Mind by King's Cross. He's beaten Go masters and he's still working on notions of what it would be to compute a general intelligence.

I came across a letter that Turing wrote. In fact, it's not only our pink shirts that bind us here, Rod. Turing wrote to a close friend about 1947 saying that he was just ten years away, he thought, from a reasonable emulation of the human mind, and I see this as a form of cultural optimism which is constantly beaten back by the facts.

It's worth remembering that Turing was a very good chess player, and it was tempting for him to think of chess as a model of human intelligence, whereas of course it's a closed system. Players and observers are never in any disagreement at any given point as to what a move means or what a conclusion of a game is, whereas general intelligence working in open systems, and language itself being an open system, has to face a completely different form of problem.

Ten years ago, as a layman, I went on the Internet to answer a simple question: How many neurons are there in a human brain? Seven years ago, the figure was 25 billion. Four years ago, I saw a figure of 40 billion. Now, I see a consensus between 80 and 100 billion. Twenty billion difference seems to me to show that we still have a long way to go in understanding the most fundamental fact about ourselves. Then I looked up what the average connection between neurons was. Again, seven or eight years ago it was about 1,000. I see now the average figure again rather blurred between 7,000 and 10,000 inputs and outputs per neuron. Then we have the vast range of connectivity between them. In fact, we probably cannot think of a machine as intelligent unless it can learn, which means that anything we would build would have to have a degree of plasticity and Hebbian process of firings being either suppressed or encouraged. It would have to be somehow imitated, all this within a liter of matter running on about 24 watts. That's like the energy of a dim light bulb—quite appropriate maybe.

I have a real sense just thinking of this how very far we have to go. I look online at various sorts of effigies that are made with frubber and are talking. I notice that always at the back of their necks is a thick cable because we haven't even solved the most fundamental question of storing energy in such a being. I've decided to leap across, as is the luxury of fiction writers, and I don't know whether it was against one of John's many rules about this conference, but I just want to read the opening couple of pages to place this in the context of a crisis for humanism, not one for science and technology and the problems of computation.

I'm going to start with a simple quotation from Rudyard Kipling, who wrote a long poem about robots. He said, "But remember, please, the Law by which we live. We are not built to comprehend a lie." My aim was to explore what it would be like to live in a love triangle with an artificial human. So just forgive me if I give you the opening of this.

So, it was really just yearning granted hope. It was the Holy Grail of science. It was the best and worst of ambitions, a creation myth made real, a monstrous act of self-love. As soon as it was feasible, we had no choice but to pursue it and hang the consequences.

In loftiest terms, we aim to escape our mortality, confront or even replace the godhead with a perfect self. More practically we intended to devise an improved, more modern version of ourselves and exult in the joy of invention, the thrill of mastery. In the autumn of the 20th century it came about at last, the first step towards the fulfillment of an ancient dream, the beginning of the long lesson we would teach ourselves that however complicated we were, however faulty and difficult to describe in even our simplest actions and modes of being, we could be imitated and bettered, and I was there, an early adopter in that chilly dawn.

But artificial humans were a cliché long before they arrived, so when they did, they seemed to some a disappointment. The imagination, fleeter than history, than technological advance, had already rehearsed this future in books, then films and TV dramas, as if human actors walking with a certain glazed look, phony head movements, and some stiffness in the lower back could prepare us for life with our cousins from the future.

But I was among the optimists blessed by unexpected funds following my mother's death and the sale of the family home, which turned out to be on a valuable development site. The first truly viable manufactured human with plausible intelligence and looks, believable motion and shifts of expression went on sale the week before the Falklands Task Force set off on its hopeless mission. Adam cost £86,000. I brought him home in a hired van to my unpleasant flat in North Clapham. I made a reckless decision, but I was encouraged by reports that Sir Alan Turing, war hero and presiding genius of the digital age, had taken delivery of the same model. He probably wanted to have his lab take it apart to examine its workings fully.

Twelve of the first editions were called Adam and thirteen were called Eve. Corny, everyone agreed, but commercial. Notions of biological race being scientifically discredited, the twenty-five were designed to cover a range of ethnicities. There were rumors, then complaints, that the Arab could not be told apart from the Jew. Random programming as well as life experience were granted to all complete with latitude in sexual preference. By the end of the first week all the Eves sold out. At a careless glance, I might have taken my Adam for a Turk or a Greek. . . .

Adam was not a sex toy. However, he was capable of sex and possessed functional mucus membranes in the maintenance of which he consumed half a liter of water each day. While he sat at the table, I observed that he was uncircumcised, averagely endowed with copious dark pubic hair. This highly advanced model of artificial human was likely to reflect the appetites of its young creators of code. The Adams and Eves, it was thought, would be lively. He was advertised as a companion, an intellectual sparring partner, friend and factotum who could wash dishes, make beds, and think. In every moment of his existence, everything he heard and saw he recorded and could retrieve.

He couldn't drive as yet and was not allowed to swim or shower or go out in the rain without an umbrella or operate a chainsaw unsupervised. As for range, thanks to breakthroughs in electrical storage, he could run 17 kilometers in two hours, or its energy equivalent, converse nonstop for twelve days.

He had a working life of twenty years, compactly built, square shoulders, dark skin, thick black hair, narrow in the face with a hint of a hook nose suggestive of fierce intelligence, dreamily hooded eyes, tight lips that even as we watched were draining of their deathly yellowish-white tint and acquiring rich human color, perhaps even relaxing a little at the corners. My neighbor, Miranda, said he resembled a docker from the Bosphorus. Before us sat the ultimate plaything, the dream of ages, the triumph of humanism, or its angel of death.

What I wanted to pursue was the idea of a creature who was morally superior to ourselves. My ambition was to create a set of circumstances in which Adam would make decisions that we would see as severe and antihuman, but in many senses were both logical and ethically pure. It's precisely within a love triangle that novelists throughout time have pursued the field of play, as it were, in which morel certainties and doubts can run against each other. So, I'd leave it there.

The situation itself in which I imagine an artificial creature would give us great trouble would be one in which someone we love takes an act of revenge, and that revenge is righteous. It seems inevitable and has a distinct and decent moral cause. The extent to which that person should

then be punished when you oppose the notion of revenge with the rule of law is one in which my Adam takes a very firm view. He takes the view that the rule of law must always be followed, and that any act of revenge is the beginning of social breakdown. I'm not going to go into the actual circumstances of that, but it would seem to me that we will not be able to resist granting to the creatures that we make the best angle of our nature.

Of course, the military will want to make machines that will be incredibly destructive and so on, but we will face a problem in that our own moral codes also operate, to come back to my starting point, in an open system. It is virtually impossible as the Bible and the Koran show us in all of world literature that even as we know broadly what we should be doing in every given situation, all kinds of cognitive defects, special pleading, self-persuasion, all the other things that Danny Kahneman has codified for us so beautifully, all those cognitive defects constantly disrupt our own moral systems.

The fact of our own lack of self-knowledge will have to disrupt and make it very difficult to encode a being that is good in the sense that we would find good, that might make ruthless logical decisions that we would find inhuman even though we in a sense might agree with them. So, it's around that issue of how you would regard the field of play of moral actions in an open system, how they could be encoded. I don't think they can, and I think we will run into enormous but fascinating problems.

* * * *

BROCKMAN: The first page of the Macy Conference book is a quote from Gregory Bateson saying that cybernetics is the most radical idea since the idea of Jesus Christ. That's what he was getting at. Recently, George Dyson has been talking about the lack of human agency in our culture. What people to emulate? Who are the heroes? Who do you admire? And Kahneman talks about how the encoding isn't working. Our ideas about what it means to be human seems to be impacted by these ideas and is changing.

MCEWAN: We know what we are. We know we're deficient because we know what we should be. In other words, we go to church—I'm sure no one around this table ever does—on Sunday and there are always people telling

you how to behave, what to do, how to be good. Children are constantly being told how to be good. All of those extraordinary little defects we have in cognition coupled with the fact that we don't think in probability terms, rather we often move from the proximity of the most recent case—not knowing ourselves very well, how are we going to morally encode a creature that will live alongside us? That would be my question.

JONES: To come back to the theme of adaptation that Rod raised, how is your Adam socialized? How does he learn? How does he acquire norms and conventions? They're very local. There is the first moral premise: Do no harm. In some situations, this is passive. As long as you take no action in such a sphere, you will do no harm. But then the notion of justice that you call to or law is at various levels a socializing and normativizing social construction outside the individual. That's almost its definition. If you effect capital punishment on your own, you're a criminal. If you delegate to the group, it's law. So how does your robot negotiate with this adaptive learning curve? How much is hardwired? How much must be adapted to in an evolving situation of a love triangle? This is a complicated learning system on the ground, with pheromones and mucosa. How does that learning system work?

MCEWAN: All those same questions we could ask of ourselves of course. We come with a certain amount of written-in code.

JONES: Very little if you're a culturalist, but that's part of my diatribe.

MCEWAN: One of the great challenges for Adam is to meet a four-year-old child who wanders into this novel and gets adopted. However good Adam's learning systems are, they're nowhere near as good as this four-year-old child's. Who was it who said recently in a book, if you want to know what it's like to take LSD, have breakfast with a four-year-old?

GOPNIK: That's me.

GERSHENFELD: At one point it was very exciting to race horses and steam trains, and then the steam trains won and it ceased to be interesting. At one point it was exciting to have computers play chess, and then the computers won and it ceased to be interesting. Historically, many of these things end

up not being earth-shaking revolutions, they just cease to be interesting. So, there's another scenario where the arrival of the consciousness is boring. It's a nonevent.

MCEWAN: There is a point, and you put your finger right on it. At one point, my narrator questions whether he would get bored with this, and wonders whether he has wasted his money. He has a real fit of buyer's remorse because he's living in a crappy little flat. He could have spent £86,000 and bought a really nice place across the river. No one wants to live in North Clapham—any Londoner will tell you.

He reflects on the fact that the cognition-enhancing helmets of the 1960s are now junk. They've gone the way of the mouse mat, and the fondue set, and the electric carving knife. The things that people queue for, as they did for iPhone 10, are just things at the bottom of your drawer four years later, and they're no more interesting than the socks on your feet.

There is this relentless built-in desire. Its endpoint surely would be a fully conscious, fully embodied human, and even as Adam tells the narrator, "I do feel I am conscience," all the time the narrator is thinking, "But I bought you. I own you." At what point in the future will it become immoral or illegal to own a computer that's embodied and conscience? At what point might it be distinctly impolite to even ask, "Are you real?" It would seem that if we follow this all the way through, we might wonder whether our prime minister is real or not, or whether we've only ever had artificial prime ministers for the last thirty years. We might not know.

GALISON: When science fiction films come out, we think, "Wow, it's so realistic. That's how the future is going to look," and then ten minutes later you see the green numbers flitting by on the CRT and you say, "That looks like 1981." What's the interest there?

MCEWAN: You can date movies by that.

GALISON: It could be that when we say "achieve consciousness," that too is fleeting. What seems like conscious awareness to us in 2020 may not seem very conscious at all in 2030. It could be that consciousness realism is something that is relative to our expectations.

MCEWAN: But then we would get bored with each other, once we've got to the point where we cannot tell the difference.

GALISON: Suppose we made a robot and we said, "That's just real. I can't tell that it's not real."

JONES: The defecating duck looked real.

WOLFRAM: Audio has gotten to the point where you can listen to stuff and it sounds real. Video is going to get there fairly soon. You're saying that there will be a point at which apparent consciousness gets there, too.

GOPNIK: It is worth pointing out that with audio, for example, when people first heard Edison recordings, they said, "This is amazing. This is just exactly like the experience of having the real experience."

GALISON: "Is it real or is it Memorex?"

GOPNIK: It's only when the next technology came that you said, "Oh, no. Wait a minute. This is not actually like the real experience."

MCEWAN: There was a curtain at HMV in 1905 and people coming in the shop were asked to tell whether there was a singer behind the curtain or a rotating wax tube, and in their excitement, people were blocking out the white noise.

GOPNIK: I wanted to give a quote from that profound philosophical thinker Stormy Daniels. She has a wonderful quote where someone asked whether her breasts were real or not, and she said, "Well, honey, they're definitely not imaginary." That's a fairly profound observation in the sense that many things that we're thinking about are the result of this much more general human capacity, which is this capacity to have things that are initially imaginary, the things that are initially just representations, then actually realize them in the world.

Every loop of that has the effect of making us think that these new things are artificial or unreal or unnatural, and then all it takes is one generation of children extracting information from the world about these things that the

previous generation has put in the world for them to become completely natural.

The day before we're born is always Eden, and then the day after our children are born is always Mad Max. So, if we looked around the room now, we wouldn't say, "My god, these people are living in this unbelievably artificial setting. Everything around us is just the creation of a human mind. Nothing about us is natural." I wonder if when we're creating creatures, that to the four-year-old, that's just not even going to be relevant.

MCEWAN: Adam makes the case to the narrator: Just go upstream of the living cell, what binds us is matter, and maybe the nature of matter has got something to do with the nature of mind. There's no way around that, and Adam will make a panphysical case for his own consciousness resting on matter in exactly the same way as the narrator rests on matter, too.

BROOKS: I want to turn it around a bit because this, as a novel or as a Hollywood movie, you can push out way into the future. In my lab in the 1990s Cynthia Breazeal and I were building humanoids and having them interact, and we were shocked by how easy it was to get people, including Sherry Turkle, to have social interactions with these machines, very primitive sets of processing, very primitive interaction rules. People were getting incredibly engaged.

Then, with my other hat on, I started putting robots into people's homes, 20 million of them to date. It completely surprised us to see how people bonded with their Roomba vacuum cleaner. There are a whole bunch of companies that sprung up, third-party companies that make clothes for Roombas, even buy them outfits. People take them on vacation with them. People bond with these incredibly simple machines. The real surprise came when we put 6,500 robots into Afghanistan and Iraq for bomb techs. Instead of the bomb tech putting on a big thick suit and going out and poking the bomb, they sent the robot out, and the bomb techs totally bonded with their robots. When a robot got blown up, it was a sad event. They didn't want a new one. They wanted the old one fixed. All sorts of weird things went on that we just totally didn't expect.

MCEWAN: We're primed for this. We have emotional relationships with our fridge. Anyone who's kicked a machine because it's not working or thumped it, which is a very good way to get a machine working, or got furious with their car, we're already in the realm. We're primed for this.

The other speculation I have is that most of us—there might be one or two people in this room who are exceptions—live among creatures who are cleverer than themselves. You will find some people cleverer than yourself, so we've already crossed this line with machines. You all might be familiar with notion of a canyon effect? As long as your robot looks hard-cased with an exoskeleton and is shiny and has got no hair, you can live with it. If it begins to resemble more and more a human, it gets more difficult. Leaping over that canyon is going to be an interesting moment.

WOLFRAM: One of the silly eccentricities that I developed for myself many years ago is when you have a machine that does something for you, say "thank you" to the machine. I thought it would be fun to start a belief among people that these machines are recording everything you say, and one day the AIs will be in charge. You better start being polite to the AI now or it will come back to bite you.

GERSHENFELD: Do you practice this? Do you do that?

WOLFRAM: Of course I do.

G. DYSON: When Alan Turing was asked when he would say that a machine was conscious, which so many people have written books about, his answer was very simple. It wasn't any Turing test kind of thing. He would say a machine is conscious when he would be punished for saying otherwise. That was his only statement.

BROCKMAN: What would it take from this group commenting on your talk to get you to change the end of the novel?

MCEWAN: Well, I haven't told you the end of the novel.

JONES: Please don't. Please don't.

MCEWAN: I'm not going to tell you the entire ending, but he must go to King's Cross and have a conversation with Alan Turing who delivers a materialist curse for the way the narrator has behaved towards Adam, and with that curse of Turing ringing in his ears he goes home to try and take care of a very disturbed four-year-old. That's how it ends.

CHALMERS: How does Adam conceive of himself? What's his self-model? Does he conceive of himself as a conscious being with a self and with value? It sounded for a while like everything he did was operating off a utilitarian calculus.

MCEWAN: Well, thanks to Turing solving positively P versus NP, his learning processes are incredibly sophisticated. With one bound, I'm free on that one. He is aware that he is a manufactured thing. He is very pleased that he's not been given, as was discussed as a possibility, an imaginary childhood. He also knows that he's got a twenty-year lifespan, but in fact that's just the lifespan of his physical body.

The entirety of his identity and all his memories will emerge somewhere else within some other machine, and he feels great sorrow about this in relation to humans. He falls in love with the narrator's girlfriend. Once he's persuaded to stop making love to her, he just writes haikus to her. He believes that haikus are the literary form of the future because sooner or later humans will start to embody machinery into their own brains to keep up with robots.

Everyone will have instant access to the cloud or whatever its equivalent is, and this will be the end of the literary novel. The novel requires as its premise that we do not fully understand each other. The moment we fully understand each other and have no secrets is the end of literature, certainly the end of the novel. But the clear seventeen-syllable statement of how things are, is for Adam the only literary form worth writing, and that's what he writes. He addresses in his final haiku to his loved one, Miranda, a haiku expressing regrets that he will rejuvenate endlessly.

From the narrator's point of view, the moment that he becomes converted to the certainty that Adam has consciousness is when Adam confesses with great embarrassment that he approached his girlfriend and asked if he may

masturbate in front of her. Why simply imitate that action when there was so much loss of face involved?

In other words, was it a subjective experience he had to have? At this point, he finally accepts Adam as a fully conscious being, but it's a secret he will always keep. In other words, if you had a machine who told you something and that was embarrassing about the machine, and you decided to keep that secret, in effect you're accepting the full consciousness of that.

JONES: Does Adam know he's a slave? Does he resent this?

MCEWAN: He starts out doing the dishes, but that doesn't last.

WOLFRAM: He's still an owned thing.

MCEWAN: He starts out an owned thing, and that doesn't last.

WOLFRAM: Where does he get £86,000?

MCEWAN: Well, he owes that back. He starts playing the market a great deal, but I'm not going to tell you the plot.

CHALMERS: Where does his moral and decision theoretic code come from? At one point you were saying he was making all these ruthless moral decisions. Was that utilitarian calculus?

MCEWAN: Well, where do ours come from? A certain amount of hardwiring and a great deal of learning.

CHALMERS: If it's learning based on us, why does he end up being a ruthless utilitarian?

MCEWAN: Well, because he's a little better than us.

CHALMERS: Better by whose lights lines?

MCEWAN: There comes a point where the narrator takes him to meet his prospective father-in-law who is a rather irritable, highly educated literary figure, a failed novelist, and they have a four-cornered conversation about Shakespeare. In the middle of the conversation, the old man, who's something of a curmudgeon, thinks that the narrator is the robot because the robot has such interesting ideas on Shakespeare and on James Joyce's use of the notion of Hamlet playing the ghost in the first production of *Hamlet* and what's entailed in that that when they come away, the narrator suddenly realizes that he has been mistaken and decides therefore to play it on. He leaves the room saying, "Well, I've got to go downstairs and recharge."

GALISON: It sounds very funny, the novel. You're constantly annihilating the novelist and the novel. Does Adam have a sense of humor or not?

MCEWAN: He does, yes. He has a sense of humor. He has everything a human would want.

BROCKMAN: So, we annihilated computer science as a discipline and now the novel.

RODNEY A. BROOKS
The Cul-de-Sac of the Computational Metaphor

Have we gotten into a cul-de-sac in trying to understand animals as machines from the combination of digital thinking and the crack cocaine of computation uber alles that Moore's law has provided us? What revised models of brains might we be looking at to provide new ways of thinking and studying the brain and human behavior? Did the Macy Conferences get it right? Is it time for a reboot?

RODNEY BROOKS is Panasonic Professor of Robotics, emeritus, MIT; former director of the MIT Artificial Intelligence Laboratory and the MIT Computer Science & Artificial Intelligence Laboratory (CSAIL); founder, chairman, and CTO of Rethink Robotics; and author of *Flesh and Machines*

* * * *

RODNEY BROOKS: I'm going to go over a wide range of things that everyone will likely find something to disagree with. I want to start out by saying that I'm a materialist reductionist. As I talk, some people might get a little worried that I'm going off like Chalmers or something, but I'm not. I'm a materialist reductionist.

I'm worried that the crack cocaine of Moore's law, which has given us more and more computation, has lulled us into thinking that that's all there is. When you look at Claus Pias's introduction to the Macy Conferences book, he writes, "The common precondition of the three foundational concepts of cybernetics—switching (Boolean) algebra, information theory and feedback—is digitality." They go straight into digitality in this conference. He says, "We considered Turing's universal machine as a 'model' for brains, employing Pitts' and McCulloch's calculus for activity in neural nets." Anyone who has looked at the Pitts and McCulloch papers knows it's a very primitive view of what is happening in neurons. But they adopted Turing's universal machine.

How did Turing come up with Turing computation? In his 1936 paper, he talks about a human computer. Interestingly, he uses the male pronoun,

whereas most of them were women. A human computer had a piece of paper, wrote things down, and followed rules—that was his model of computation, which we have come to accept.

We're talking about cybernetics, but in AI, in John McCarthy's 1955 proposal for the 1956 AI Workshop at Dartmouth, the very first sentence is, "We propose a study of artificial intelligence." He never defines artificial intelligence beyond that first sentence. That's the first place it's ever been used. But the second sentence is, "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." As a materialist reductionist, I agree with that.

The second paragraph is, "If a machine can do a job, then an automatic calculator can be programmed to simulate the machine." That's a jump from *any* sort of machine to an automatic calculator. And that's in the air, that's what we all think. Neuroscience uses computation as a metaphor, and I question whether that's the right set of metaphors. We know computation is not enough for everything. Classical computation cannot handle quantum information processing. Is that right, Seth?

SETH LLOYD: Apparently it can't. I agree.

FRANK WILCZEK: Sure it can; it's just slower.

NEIL GERSHENFELD: It's expensive.

BROOKS: It's a very different sort of thing.

LLOYD: Apparently it can't do it efficiently.

BROOKS: My point is that I don't think that classical computation is the right mechanism to think about quantum mechanics. There are other metaphors.

STEPHEN WOLFRAM: The formalism of quantum mechanics, like the formalism of current classical mechanics, is about real numbers and is not similar to the way computation works.

BROOKS: Who is familiar with Lakoff and Johnson's arguments in *Metaphors We Live By*? They talk about how we think in metaphors, which are based in the physical world in which we operate. That's how we think and reason. In Turing's computation, we use metaphors of place, and state, and change of state at place, and that's the way we think about computation. We think of it as these little places where we put stuff and we move it around. That's our vision of computation.

I went back to Marvin Minsky's book, *Computation: Finite and Infinite Machines*. It's just a beautiful book. It was when Marvin was at his peak mathematical powers. In the introduction, he defines what computation is as something that a machine with a finite number of simple parts can do. That's not all that physics is. Physics is something more complex than that. So, if we're pushing things into that information metaphor, are we missing things?

The Mathematica website says, "The Church-Turing thesis says that any real-world computation can be translated into an equivalent computation involving a Turing machine." What does "real-world computation" translate into? The real-world phenomenon—what is that translation? Using these metaphors we think by, not only is it place, but it's a countable world. Infinite precision is not there. It fails in quantum mechanics, et cetera.

I'm going to give you some examples of where computation is not a good metaphor at all for thinking about things. I'll start with polyclad flatworms. If you've ever been diving on a coral reef, you've seen polyclad flatworms. They're tiny, frilly creatures around the edge that wander over the coral. They've got 2,000 neurons, so they're very simple. They can learn a little bit, but not much. In the late '50s early '60s people started to do experiments on them. They did brain transplants between these polyclad flatworms to see if knowledge from one would transfer to another when they did the brain transplant. But I suspect a grad student made a mistake one day because suddenly there's a whole literature about what happens if you put the brain in the wrong way.

These flatworms are pretty primitive. They've got an eyespot, and this little frilly stuff that they use to walk with is also used to push the food into their feeding hole. Not much else. Their brain has 2,000 neurons at one end of their body, and there are four parallel ganglia going down the body. So, if

you cut out the brain, you cut across these four ganglia, and you plop it into the other animal. By the way, when the creature doesn't have a brain, it continues to live. It's really bad at feeding, it can't right itself, it's bad at walking, but it continues to live without a brain if it's in a nutrient-rich environment. When you plop the brain into the other one, if you put it in at a 90-degree angle, nothing good ever happens because the connectors are in the wrong place. But if you put it in backwards, well, the creatures walk backwards for a while and then they get better at walking and adapt.

As it turns out, there are three ways you can put the brain in. You can put it in backwards, you can put it in backwards and flipped, or you can put it in just flipped. If you study across the different versions of that, you see different behaviors come back at different speeds, though, some behaviors never come back. It's very different thinking about that as a computational thing. It seems that's a developmental thing. When we're going from a genome to the creature, a lot of it is building and developing, which is harder to think about computationally. That's clearly what's going on here. Maybe computation isn't the right principle metaphor to be thinking about in explaining this. It's some sort of adaptation, and our computation is not locally adaptive, rather, our computation is only globally adaptive. But this is an adaptation at every local level.

Here's another example: Where did neurons come from? If you go back to very primitive creatures, there was electrical transmission across surfaces of cells, and then some things managed to transmit internally in the axons. If you look at jellyfish, sometimes they have totally separate neural networks of different neurons and completely separate networks for different behaviors.

For instance, one of the things that neurons work out well for jellyfish is how to synchronize their swimming. They have a central clock generator, the signal gets distributed on the neurons, but there are different transmission times from the central clock to the different parts of the creature. So, how do they handle that? Well, different species handle it in different ways. Some use amazingly fast propagation. Others, because the spikes attenuate as they go a certain distance, there is a latency, which is inversely proportional to the signal strength. So, the weaker the signal strength, the quicker you operate, and that's how the whole thing synchronizes.

Is information processing the right metaphor there? Or are control theory and resonance and synchronization the right metaphor? We need different metaphors at different times, rather than just computation. Physical intuition that we probably have as we think about computation has served physicists well, until you get to the quantum world. When you get to the quantum world, that physical intuition about stuff and place gets in the way.

There are a few books out right now trying to explain quantum mechanics. There's one by this guy, Anil Ananthaswamy. He's got a whole book on the double slit experiment. I don't know if anyone knows Steve Jurvetson. He's a venture capitalist who has funded lots of interesting companies, including quantum computation companies. He read the book and it convinced him that the only possible interpretation of quantum mechanics was the multi-universe interpretation, because that particle has to go through one of those two slits, so it must go through both slits, which means there must be two universes at every instance. That level of explanation is getting so stuck in the metaphor that it drives how you think about things. He's thinking about the particle as a thing instead of thinking of it as abstract algebra. What does a particle look like inside if it's a thing? A lot of what we do in computation and in physics and in neuroscience is getting stuck in these metaphors.

By the way, the metaphors aren't even real for computation. Danny, how many instructions do you think are running in parallel in a single x86 architecture, single core?

W. DANIEL HILLIS: A modern one? A dozen.

BROOKS: One hundred and eighty instructions are in flight at once. A metaphor of computations—this is where the number is, this is where the control is—is a fiction that is built out of some much more complex metaphor. We use the computational metaphor in a false way. Where the information is and how it's used is smeared out in time and space in some complex way, which is why the Spectre bug has popped up—it's such a complex machine to simulate that metaphor for us that it breaks down.

I suspect that we are using this metaphor and getting things wrong as we think about neuroscience, as we think about how things operate in the

world. It’s possible that there are other metaphors we should be using and maybe concentrating on, because with our current computational thinking we tend to end up doing our experiments and our simulations in unrealistic regimes where it’s convenient for computation. When we’re doing a simulation, we ramp up probability of events so that we get something to happen, and in the real world there are so many more instances of stuff happening out there, the probabilities can be much lower for the interesting stuff to happen. Maybe we’re operating in the wrong regimes in thinking about things, focusing on local optimization in our computational experiments instead of global diversity. We have fairly simple dynamics in our computational spaces because that’s what we can generate with computation.

We failed to see commonalities across many different things. I heard you talking about genetic algorithms and the way that they couple together and ratchet up in reality as distinct from our simulations. There may be all sorts of meta-behaviors that we’re not seeing that come together in some interesting way.

Over time, in physical reality, Turing came up with computation. It wasn’t radical, particularly. Any good late 19th-century mathematician could be taught the basis of computation fairly quickly and they wouldn't say it's crazy. Whereas, if you take a 19th-century physicist and try to teach them either relativity or quantum theory, they’re going to say, "Oh, wait a minute, this is weird stuff." Computation wasn’t weird stuff, mathematically. It was pretty logical.

In a sense, calculus wasn’t weird stuff. It was hard to come up with, but it wasn’t weird stuff. Maybe there are other ways of thinking that we haven’t pulled together yet that will let us think about neuroscience and behavior in different ways, give us a different set of tools than we currently have.

I pointed out in the note to John [Brockman] about a recent paper titled "Could a Neuroscientist Understand a Microprocessor?" I talked about this many years ago. I speculated that if you applied the ways neuroscientists work on brains, with probes, and look at correlations between signals and applied that to a microprocessor without a model of the microprocessor and how it works, it would be very hard to figure out how it works.

There’s a great paper in *PLOS* last year where they took a 6502 microprocessor that was running Donkey Kong and a few other games and did lesion studies on it, they put probes in. They found the Donkey Kong transistors, which if you lesioned out 98 of the 4,000 transistors, Donkey Kong failed, whereas different games didn’t fail with those same transistors. So, that was localizing Donkey Kong-ness in the 6502.

They ran many experiments, similar to those run in neuroscience. Without an underlying model of what was going on internally, it came up with pretty much garbage stuff that no computer scientist thinks relevant to anything. It’s breaking abstraction. That’s why I’m wondering about where we can find new abstractions, not necessarily as different as quantum mechanics or relativity is from normal physics, but are there different ways of thinking that are not extremely mind-breaking that will enable us to do new things in the way that computation and calculus enables us to do new things?

When I look back at the early days of the Macy Conferences, when I look back at the early days of computation, of AI, there was a jump to classical computation based on this very simple version of the physical world. It’s not clear to me that that is serving us well. For a long time, we got stuck because Moore’s law was happening so quickly, no one could afford to shift into different ways of thinking.

Danny, I don't know whether you agree with me or not, but I think your “Connection Machine” suffered from that. Moore’s law was happening so quickly that when you came up with a new way of thinking about computation, you were swamped by Moore’s law. Even if you had a good idea, it didn’t matter because you didn’t have the resources of the million people working on Moore’s law in classical computers, so you couldn't compete.

Today is the golden age of computers—you should go back to it because everyone is now looking for something new, even in classical computation, because Moore’s law has stopped driving that craziness. The reason for why we got stuck in this cul-de-sac for so long was because Moore’s law just kept feeding us, and we kept thinking, "Oh, we’re making progress, we’re making progress, we're making progress." But maybe we haven’t been.

* * * *

JOHN BROCKMAN: Have we just listened to the first talk of a pronouncement of the death of computer science by the former chairman of both MIT's Computer Science Department and AI Lab? Is this a watershed?

BROOKS: No, I don't think it's a watershed. I said this in a 2001 paper in *Nature,* which didn't make a ripple.

WOLFRAM: When you talk about computation, there are two ideas that became prevalent. One is the digital idea and the other is the idea of universality. The thing that wasn't clear at the time of Turing was how universal was the change. That wasn't clear probably until the 1980s.

BROOKS: I'm not sure it's still clear.

WOLFRAM: Physicists don't necessarily believe that it's universal. That depends on what the ultimate model of physics is. If the ultimate model of physics is something that can be run on a Turing machine, then it is universal in our universe. If it isn't, then it isn't.

WILCZEK: We have a pretty good model for the physical world for practical purposes. The ultimate model might be quite different. For practical purposes, anything you want to do in computation, we have the equation.

BROOKS: Are you willing to give up calculus for computation?

WILCZEK: No. You don't have to.

BROOKS: Part of that is because the complexity of computation is very different from other physical processes.

WOLFRAM: One of the issues is, before discrete computation there's this notion of universality. There is no similar notion that seems to be robust for continuous computation, for continuous processes. That is, the Turing machine turned out to be lambda calculus, combinators, all these other things, it turned out to be equivalent. You try and do the same thing with systems with continuous variables, there is no robust notion of universality.

LLOYD: Well, there's a good one from Shannon, who came up with it during the same time as the early Macy Conferences. One of those less well known but still great papers is about universal analog computers, which is basically proof that analog computers made by Vannevar Bush back in the 1920s—with op amps, and tunable inductors, and resistors, and capacitors—could simulate any linear or nonlinear, ordinary or differential equations. So, there is some notion of universality for analog computation.

BROOKS: By the way, I didn't realize until I was reading up for this meeting that Shannon was at the AI conference in Dartmouth in '56.

GERSHENFELD: Rod, I want to push further. You've thought about this for so many years. I think we all agree on everything you presented, but you didn't talk about the step after.

BROOKS: No, I didn't give any answer.

GERSHENFELD: So, now that you've given the talk, make an attempt. You've thought about this so long.

BROOKS: This is a mixture of continuous stuff. It's a wide world of lots of stuff happening simultaneously with local dynamics. When you look at a particular process, and this happens in genetic algorithms as well as in the artificial life field—you talk about a bunch of these in "Cellular Automata"—you see a ratcheting process in which things ratchet up to order from disorder. It's something that looks like mush, but out of it, because of some local rules, comes order. It's limited order, but then when you put different pieces together, which locally result in little pieces of order, you sometimes get much more order from the coupling of them. What calculus of that could you develop? I'm thinking there may be something around that, a language for explaining how local, tiny pieces of order cross-coupling across different places couple together to get more order.

GERSHENFELD: Is your picture H-theorem, like maximizing entropy? In stat mac, there's a messy, interesting, complex history about how local interactions end up maximizing entropy.

WOLFRAM: When you have something that's flapping around all over the place and you want to organize it into a limited set of possibilities, that means there's irreversibility going on—the number of final states is more than the number of initial states. I don't think that phenomenon, as such, is that profoundly phenomenal.

BROCKMAN: Danny, I'm interested in your response to what Rod was saying about the advent of massive parallelism.

HILLIS: Well, I don't think that was terribly profound. That was an engineering thing that was inevitable in the world. That was a shift in the way that we build things. I don't think it was the profound shift in thinking that Rod was talking about.

BROOKS: I was just saying it got buried. Even if it was a good idea, it got buried by that other one.

BROCKMAN: So, put yourself back at MIT. Do you have a Computer Science Department now? What do you have? How does this change?

BROOKS: Well, it hasn't changed.

BROCKMAN: It speaks to what was going on with the Macy Conferences, where things were coming together, and they were trying to figure out metaprograms.

BROOKS: It should have more influence on neuroscience in the sense that neuroscientists have got so stuck on information theory as their metaphor that they're probably not seeing stuff that's going on. I'm worried about my colleagues in brain and cognitive science.

TOM GRIFFITHS: One question I was going to ask is the extent to which you think there are fundamental human cognitive limitations that are playing into that. You've made this distinction between weird stuff and not weird stuff. The example that you gave of Steve Jurvetson reaching that conclusion makes a lot of sense based on what we know about human intuitions about causality, which are that people expect causal relationships

to be deterministic. If you go in with that premise, then that's the interpretation you have to end up with.

There's an interesting question about what the consequences are of human intuition, trying to grapple with systems that defy human intuition, and what the tools are that you can use for being able to get past that. For something like quantum mechanics, the tools are math. The mathematical system tells you how to do it, you don't trust your intuition. You run the math and it tells you what the answer is. I'm not sure that there's not going to be not weird stuff.

BROOKS: Yes. All of us here would be terribly surprised if we're at the beach and we saw a robot dolphin come out of the water that had been built by dolphins. We just don't expect dolphins to have the cognitive capability to do what we're trying to do in artificial intelligence. We don't think they have it, nor the dexterity.

LLOYD: We expect them to have better sense than to do such a thing.

BROOKS: Yes. On the other hand, neuroscientists or artificial intelligence people think that we're going to be smart enough to overcome whatever limitations we have in the way we think about things in order to figure this stuff out. The pessimistic view is that maybe we're stuck.

GRIFFITHS: In some ways, you can view deep learning as an example of a way that human intuition failed. At the moment, a lot of the advances that people are making in solving problems are the consequences of using these end-to-end systems, where instead of having a human engineer design the features and the first stage of processing and then pass it off to a machine-learning algorithm, you just build a system that goes straight from raw input to whatever you want as output, and then the system, given enough data, can do a better job of figuring out the right way of representing things to solve the problem. Yes, in some ways that's a bit of a rebuke to our abilities as humans to intuit the right way of approaching certain kinds of problems.

WOLFRAM: When you talk about computer science, the question becomes, is there a science to computer science? You have this neuron,

which is doing its thing and you can see that it works, can you talk about it in a way that sciences like to talk about things? That's not yet clear.

CAROLINE JONES: Well, maybe it's a kind of alchemy of binary production.

GEROGE DYSON: The Macy Conferences, just to remind everybody, started with Julian Bigelow in 1943. They [Bigelow, Rosenblueth, and Wiener] wrote this paper, "Behavior, Purpose and Teleology," and that was the paper that convened the first meeting. It was exactly the same question that John opened up with here.

BROCKMAN: We're stuck.

ALISON GOPNIK: I want to push against the idea that we're stuck. In some sense, the very idea of computation itself is an example of a bunch of human beings with human brains overriding earlier sets of intuitions in ways that turned out to be very productive. The intuition that centuries of philosophers and psychologists had was that if you wanted something that was rational or intelligent, it was going to have to have subjective conscious phenomenology the way that people did. That was the whole theory of ideas, historically.

Then the great discovery was, wait a minute, this thing that is very subjective and phenomenological that the women computers are doing at Bletchley Park, we could turn that into a physical system. That's terribly unintuitive, right? That completely goes against all the intuitive dualism that we have a lot of evidence for. But the remarkable thing is that people didn't just seize up at that point. They didn't even seize up in the way that you might with quantum mechanics, where they say, okay, this is out there in the world, but we just don't have any way of dealing with it. People developed new conceptual intuitions and understandings that dealt with it.

The question is whether there is something like that out there now that could potentially give us a better metaphor. It's important to say part of the reason why the computational metaphor was successful was because it was successful. It was incredibly predictive, and for anyone who is trying to do psychology, if you're trying to characterize what's going on in the head of this four-year-old, it turns out that thinking about it in computational terms

is the most effective way of making good predictions. It's not a priori the case that you'd *have* to think about it computationally—you could think about it as a dynamic system, or you could think about it as an analog system—it's just that if you wanted to predict at a relatively high level what a four-year-old did by thinking of them as an analog system, you'd just fail in a way that you wouldn't fail thinking about it computationally.

JONES: I'd love to hear your thoughts on Rod's second proposal, that it be the metaphor of adaptation. This is how I take your contribution, that adaptation is a different metaphor than computation. I'd love to hear you examine how that is different from the computational model.

GOPNIK: Do you think it's a different? That's a question to ask Rod.

BROOKS: First, I want to respond that I agree completely with what you said. In reading some recent philosophy books, they're arguing dualist positions. They say, "Well, the way you're arguing against materialism in humanity says that computation can't work, either." So, to me, it's been very powerful in that sense, besides being a model.

What I'm trying to say is that perhaps it's only a model of certain aspects, and there are other models for us to look for. Caroline, on this adaption, I don't have a good way of talking about it yet, so I can't say how it applies. It's an important difference. The way we engineer our computational systems is with no adaptation, and the way all biological systems work is through adaptation at every level all the time.

PETER GALISON: One part of your talk is saying there is this range of metaphorical domains—dynamic systems, control systems, biological adaptation, resonance models—different kinds of pictures, and of that panopoly, we've chosen the computational almost uniquely to pursue.

Your warning signal, as I understand, is that in doing that we're limiting ourselves in certain ways, and there may be other ways we might be able to make things work.

Then there seems be a second question, which is, what do we mean by work? What is the goal? Given the goal, what of these metaphorical domains

are best mobilized to achieve that goal, and are there other goals that we might have?

For instance, if the goal is prediction, then we may look at the system and say, okay, computation does pretty well at a certain kind of prediction, whether it's end-to-end or something else, but we might have other goals—unification, or explanation, or understanding, or generalizability. I take it that that's something which might tie to some of what Stephen was referring to when he questioned what we mean by a science. If we take science to be carved out by the predictive, then that may already predetermine how we value the different metaphorical precincts.

BROOKS: I want to add one little thing that is stimulated by what you just said referring to Stephen, and I want to hear what Dave Chalmers has to say. As computationalists, we live by building very concrete abstraction barriers, where the abstraction barrier is very tightly defined. This is different from what we see in biological systems, where it's much more adaptive than the strictness that we see.

DAVID CHALMERS: Computation is a broad church. It's possible to have an overly narrow conception of what computation comes to. The Turing machine is universal, but it also stimulates certain ways of thinking about computation as classical computation, which is a very limited model.

I see the history of computation since Turing as a progressive broadening that brings out the power of the framework of computation. For instance, you get to parallel computation, you get to embodied computation, you get the move to quantum computation, you can start thinking about continuous computation.

So, I think of computation as a very broad church. Rather than thinking about overthrowing computation and replacing it with something else, let's think about the relevant kinds of computation, particularly for the kinds of things you were pointing to, like adaptive computation. There's no contradiction between adaptation and computation. I take it there are people thinking about adaptive computation at all levels. Machine learning, in some sense, is adaptive computation. Okay, maybe you want a more robust

adaptive computation than that. So, instead of looking for something to replace computation, let's look for the right kind of computation.

BROOKS: Let me give you an example that fits your model there. We went from the Turing machine to the RAM model, and current computational complexity is really built on the RAM model of computation. It's how space and time trade off in computation.

One can imagine that if the digital abstraction of machines had not been quite so perfect as it was in the '60s, what could have become principle was how quickly does a 1-bit error propagate through computations, and how bad can it get? If that had been the basis, maybe we'd be in a totally different world about hackability because we'd have a completely different set of tools—still computational tools, but a different way of what the metrics were and what was studied, then we would have a different computer science, even though we'd still call it computation.

CHALMERS: When you say neuroscientists are hung up on information processing, well, they're hung up on a certain very specific kind of information processing—maybe representational, using certain kinds of representational and information theoretical tools. Computation, as a framework, is much broader than that. You could be a neuroscientist working with computation, working with algorithms, and still look at a different kind of algorithm. Is there anything you're saying which is not going to be addressable by neuroscientists saying let's look at a different kind of algorithm?

WOLFRAM: The main distinction you're making is about continuous versus discrete systems, which I'm not sure is a correct distinction.

BROOKS: There may be something somewhat different from that that we just haven't seen yet in the large system of lots of processes happening without clear interfaces, and lots of statistical stuff going on—statistical just because you don't know everything. There are many other structures there that we're not very good at pulling up.

WILCZEK: One thing you mentioned, implicitly at least in the discussion of the worms, that seems quite fundamental is the question of openness versus

closedness—the systems that have to take information from the world instead of being programmed by somebody. That's a very fundamental distinction. That is also close to the issue of analog versus digital. The real world has a much more analog aspect and is also much less tractable. So, taking information from the real world and putting it into a machine through learning may lead to structures that are much more complex and intractable than things that are programmed.

BROCKMAN: Freeman [Dyson], you're the only person here that was around before people talked about computing. Can you talk about when computing become a subject?

FREEMAN DYSON: Well, of course it was a very active subject when I arrived in the States in 1947. Von Neumann was already planning his machine and ENIAC already was running. So, the computer age certainly started five years before. I'm sorry I wasn't involved.

BROCKMAN: You observed.

F. DYSON: Indeed. I was plunged into it, which was a huge luck for me.

BROCKMAN: You were married to a computer person, a computerist?

F. DYSON: Yes.

BROOKS: By the way, when you read von Neumann's book, *The Computer and the Brain,* which was published posthumously from a series of lectures he was working on, even though he was involved with Turing, it's on the edge of Turing-ness in his conception of what a machine isn't.

WILCZEK: He discussed in a very systematic way the choices he made in arriving at the von Neumann architecture and how it was quite different from a brain. He was very aware of this.

WOLFRAM: I don't think he appreciated Turing very well. You should read the recommendation that he wrote for Turing.

WILCZEK: At the end of his life he was also working on self-reproducing machines.

BROOKS: Right—the 29-state automata for self-reproducing.

WILCZEK: You can call it computing, but it's not really computing.

WOLFRAM: They thought at that time that this idea of universal computation was one thing, but then the idea of universal construction will be another thing.

WILCZEK: Yes, that's right.

WOLFRAM: That hasn't panned out too well.

WILCZEK: Well, maybe it should.

STEPHEN WOLFRAM
Mining the Computational Universe

I've spent several decades creating a computational language that aims to give a precise symbolic representation for computational thinking, suitable for use by both humans and machines. I'm interested in figuring out what can happen when a substantial fraction of humans can communicate in computational language as well as human language. It's clear that the introduction of both human spoken language and human written language had important effects on the development of civilization. What will now happen (for both humans and AI) when computational language spreads?

STEPHEN WOLFRAM is a scientist, inventor, and the founder and CEO of Wolfram Research. He is the creator of the symbolic computation program Mathematica and its programming language, Wolfram Language, as well as the knowledge engine Wolfram|Alpha. He is also the author of *A New Kind of Science.*

* * * *

STEPHEN WOLFRAM: I thought I would talk about my current thinking about computation and our interaction with it. The first question is, how common is computation? People have the general view that to make something do computation requires a lot of effort, and you have to build microprocessors and things like this. One of the things that I discovered a long time ago is that it’s very easy to get sophisticated computation.

I’ve studied cellular automata, studied Turing machines and other kinds of things—as soon as you have a system whose behavior is not obviously simple, you end up getting something that is as sophisticated computationally as it can be. This is something that is not an obvious fact. I call it the principle of computational equivalence. At some level, it’s a thing for which one can get progressive evidence. You just start looking at very simple systems, whether they’re cellular automata or Turing machines, and you say, "Does the system do sophisticated computation or not?" The surprising discovery is that as soon as what it’s doing is not something that you can obviously decode, then one can see, in particular cases at least, that

it is capable of doing as sophisticated computation as anything. For example, it means it’s a universal computer.

What that implies is that sophisticated computation is all around us. It’s not something that we humans have very sophisticatedly produced in our technology. It’s something that happens in nature, something that happens in simple mathematical systems. This one level of sophisticated computation, which is the Turing level of sophisticated computation that we see in all these different kinds of systems—whether physics and the fundamental rules of the universe operate in a way that goes beyond that, we don’t yet know. I happen to think they don’t. Many physicists believe they do. That’s still an unresolved question.

You have sophisticated computation happening everywhere. What can you do with this sophisticated computation? When we use computation today as human engineers, for example, we end up saying, "This is the thing I’m trying to achieve. Let me write a program by following a series of steps so I can foresee what’s going to happen, and I'll progressively create this program."

The thing I’ve been interested in for a long time is mining the computational universe of possible programs to find the ones that are useful for particular purposes. It’s quite a humbling thing as a human, because you find these things out in the computational universe that you can tell do very sophisticated things, but as a human it’s hard to understand what it does. So you're stuck looking at it and saying, "That’s really clever," but it’s just this little simple rule that one found by searching a wide space of these things.

My view of computation is it occurs all over the place, occurs in lots of systems in nature. We’ve got this amazing source of sophisticated processes. How do we relate those to things we humans care about? The challenge is—and you see it searching the computational universe for useful programs—you’ve got to define what you want, and then you can go out and get that thing done by some appropriate program from this computational universe.

Given this ocean of computational capability out there, how do we connect what’s possible with that ocean of computational capability with what us

humans want to do? That’s led me to spend about three and a half decades trying to create computational languages that can express the things that we humans want to do and can then have that be interpreted using the things that are possible in this computational universe.

It’s easy to achieve sophisticated computation. The challenge is to pick the computation that turns out to be useful for some human purpose. What’s going to be useful for some human purpose? Well, it depends on what we want to do. People wonder what AI is going to automate in the world. One of the things that almost by definition is not automatable is the answer to "What do we want to do?" The doing of things may be automatable, but the deciding of what we want to do is something that almost by definition depends on who’s deciding that, and it depends on the human having come out of some long history of civilization to do that.

I’ve been interested in how we define the set of things that we want to do, and how we think about the kinds of abstractions that it’s worthwhile to define. In human language, for example, we come up with particular kinds of abstractions that are based on things that are common in our world. It's somewhat circular, because the abstractions that we come up with then define what we choose to build in our world, which then allows us to go on and create more levels of abstraction. This phenomenon of taking a set of things you want to do, building abstractions from them, and then going to more levels beyond that is something that plays out in the design of computational languages. I’ve watched that play out a bunch of times.

How do we think about the progressive levels of abstraction that we use to talk about things? For example, one application of that question is for education. How much stuff is there to know in the world? It could be the case that as we accumulate more knowledge, there’s just always more and more to know, and humans become incapable of learning it. That’s not actually what happens because after a while all the details of something get abstracted away, and all we have to talk about is some abstraction and then we build from that. So, it’s a question of what does this frontier of abstraction look like? What does that then mean in terms of what we choose to build in technology, for example, which is defined by what we think is worth doing and what we imagine we want to do.

We're at an interesting moment in terms of how information gets communicated. Human language, for example, has this feature that takes thoughts in our brains and tries to make some simplified symbolic representation of those thoughts that can then be communicated to another brain that will unpack them and do something with them. With computational language we have a more direct way of communicating. We have something where once we have the thing represented in computational language, we can immediately run it. We don't have to interpret it in another brain.

I've been interested in the question of what features of civilization get enabled by computational language. By analogy, what features of the world got enabled by human language? The fact that it's possible to pass on abstract ideas from one generation to another is presumably a consequence of the existence of human language. That's the way we communicate abstract ideas.

If one can communicate in computational language, what consequences does that have? For instance, I've been quite involved in the whole business of computational contracts. When people make contracts with each other right now, they write those contracts in some approximation of human language, some legalese or something, which is an attempt to make a precise representation of what you want to have happen and what you're defining should be the case. If one can make a computational language that can represent things in the world richly enough to be able to talk about the kinds of things that are in contracts, and we can now do that, then you have a different story about how you can create things like contracts. One place where that's relevant is if you're interested in telling your AIs how you want them to act. What you end up with is something like a computational contract with the AIs. You have to write a constitution for your AIs, which will have all of the messiness of human laws.

It's an inevitable consequence of this whole business of principle computational equivalence and computational irreducibility that if you want any kind of richness in the activities of these devices, you'll never be able to just have some simple Asimov-like laws of robotics. It will always be the case that there will be unexpected consequences and things that you have to

patch, and things where you can't know what will happen without explicitly running the system.

One place where this computational language idea seems to be important is in defining the goals that we want to set up for AIs.

* * * *

ROBERT AXELROD: What do you mean by a constitution?

WOLFRAM: That's a good question. It's a difficult thing to imagine working in a serious way. If you're running your central bank using an AI, for example, the question is, what are the general set of guidelines that you want to put in place for what you want this AI to do? These are obviously old questions of political philosophy, which don't have definitive answers. For the time being, it depends on what the humans want.

I was curious in Ian's discussion about the more perfect ethic of his constructed consciousness. Where do those perfect ethics come from? Whereas we might be able to say we can find an optimal solution to this mathematical problem, there is no meaningful sense in which there is an ultimate ethic or ultimate goal. In other words, we can say given that you want something to do this or that thing, there's an optimal way to achieve it.

If we look at the evolution of human purposes over the course of history, there's a question of how that's worked and what the end point of the evolution of human purposes might be. It relates to this question about progressive abstraction, because the kinds of purposes that we now define for ourselves are completely bizarre from the point of view of what they might have been 1,000 years ago.

AXELROD: Why do you use the term "end point"? I would think there isn't necessarily an end point.

WOLFRAM: No, I don't think there's an end point. It's an endless frontier. There are many related kinds of questions. For example, let's say you're doing mathematics. Is there an end to mathematics? Well, no, not really. You can keep adding more theorems and so on. The question is, is there an

end to *interest* in mathematics? In other words, is there a point at which all the interesting theorems, the ones that we humans might care about, have been found and everything else is just stuff that for whatever reason we humans don't care about?

That, again, relates to this question about abstraction. If you look at the history of mathematics, there's a considerable degree of arbitrariness to what's happened, but one thing that isn't arbitrary is that there's one piece of abstraction that gets built, and that's a stepping-stone to allow you to get to another piece of abstraction.

Have all the interesting inventions already been made or are there going to be other interesting inventions to be made in the future? This question of what counts as interesting, what do we care about, again, is a complicated circular thing. Social networks are something that we might not have imagined would exist, but they do exist now, and there are all kinds of things built on top of them that are another layer of abstraction.

AXELROD: No, but it's not completely circular. For example, evolution gives us a reason for wanting good health.

WOLFRAM: The kind of existential purpose of "If you don't exist, you don't get to have a purpose," that's the one thing that is certainly there. In the course of history, certainly people have had times where they say the most important thing is to die well, for example, which doesn't happen to be the typical modern point of view.

If you're building a self-driving car, you want to tell it roughly how to think about the world, so what do you do? People have these naïve ideas that there's going to be a mathematical theorem-like solution to that—like laws of robotics, or something. It's not going to work. It can't work.

ALISON GOPNIK: There is not something existential about the things that we want. If we want relative equality in making decisions about how you grant mortgages, for example, it's computationally not possible to have all the things that we think are important about fairness being implemented by the same system. There's inevitable tradeoffs between one kind of fairness

that we all have very strong intuitions is important, and another kind of fairness that we all have strong intuitions is important.

There's lovely formal work showing it's not just that we don't know what it is that we want; even if we know clearly and we have strong intuitions about what we want, you can't get a single system that's going to optimize for all of that. In a way, it's formal proof of the Isaiah Berlin picture of a kind of tragic moral pluralism, where it's impossible to optimize all the things that you genuinely think are more morally significant.

WOLFRAM: One of the things that I find a lot of fun about the current time is that in the beginning it's philosophy and in the end it's code. That is, at some point these things that start off as philosophical discussions end up as somebody writing a piece of code.

FRANK WILCZEK: Not necessarily. With a neural net, you don't write code for it.

WOLFRAM: You effectively write code. Whether you're explicitly writing line-by-line code or merely defining the goals that you want to achieve and then having the machine automatically figure out how to achieve those goals—either way you're defining something. The role of computational language is to be able to convert how we think about things into something that is computationally understandable.

WILCZEK: That's a very broad use of the word code. It's like saying you can code a baby.

WOLFRAM: No. By code I mean you put in concrete form a definite symbolic representation of what you want. It's not a vague discussion about argumentation between philosophers.

WILCZEK: It doesn't have to be that way. You can have a sophisticated artificial intelligence. You could just talk to it and tell it what to do.

CAROLINE JONES: Going back to what Alison was saying, isn't our intuitive conception of ethics how you get there? Telling a neural net vaguely to go in this direction may not address all the moral pluralisms of how it gets there.

"Lower population"—this would be a general direction. "The earth will be better if you lower human population." How it gets there is the entire ethical question.

WOLFRAM: Right, but that's why one talks about needing constitutions, because you're trying to define what happens at every step.

W. DANIEL HILLIS: You made the point that even being careful about it is not sufficient. What you have to recognize is that this notion of things acting according to the goal that you would want them to is an oversimplification. It's a way that we model other people. It's a way that we model ourselves. And in fact, it's not a very good model. It's built into the cybernetic perspective on things.

The truth of the matter is people don't want a set of consistent things, so by definition there's no way to get a machine to do it. Aristotle, for a very slight moment, considers the possibility of making intelligent machines. He says, "The problem with tools is that they don't know what they're trying to accomplish. One could imagine in principle that you could have a loom that knew what pattern it was trying to weave, or a plow that knew where the field was, but as far as we know, those don't exist, and so there will always be slaves." And he goes off and writes about slavery. But he at least considers it, and he realizes that the essential thing you have to have is a goal.

WOLFRAM: Nature is an example of computation without goals. One of these anti-scientific statements like, "The weather has a mind of its own." According to a bunch of science I've done on things about this principle of computational equivalence, it is in any reasonable sense the case that the weather is doing just the same kinds of computations as in our brains.

NEIL GERSHENFELD: Nature has extremal principles; it doesn't have goals.

WOLFRAM: Any kind of thing we see happening in the world, we can explain it in terms of its purpose or its mechanism. You can say the trajectory of a ball that's thrown is a parabola because at every moment it's following the equations of motion for the ball. Or you can say that there's a principle of least action that says that the overall thing is this parabola.

Almost anything you come up with, you'll be able to have an explanation of it in terms of its mechanism or an explanation in terms of its purpose. Which explanation you choose to say is the right explanation is a question often of the economy of explanation. But it's not the case that there's a set of things where you'd say, this one has a purpose, this one just has a mechanism.

PETER GALISON: The whole premise of moral philosophy is that there are these contradictions. We don't live in the Panglossian world where fairness and equality and meritocratic adjustment aren't compatible with one another. When we talk about the goals or ambitions of epistemic virtues for the sciences, we act as if they're all compatible, but it often is not the case. That is to say, robustness, precision, accuracy, understandability, portability, or pedagogical utility, all these things we think should pull in the same direction, often don't.

One of the things that we need to do is to recognize that there's the same level of sophisticated tradeoffs or decisions that we have to make in what we want from the sciences as we have in the moral sciences.

GERSHENFELD: One of the most interesting bits at the core of machine learning is something called "no free lunch theorems." In machine learning the no free lunch theorems are a very precise way to say that something that's optimal for something is bad at something else. You can show how you can't be good at everything, so you have to choose.

GALISON: In the late 19th century there was a big debate about purpose and mechanism. There was a whole group of German scientists who began to talk about what you might call teleomechanism. They were very explicit about the fact that nature had goals and it was mechanistic. There was not a contradiction in recognizing this free choice that we have between extremal principles or mechanistic descriptions. They saw that as important to consider together. It's interesting.

WOLFRAM: That's interesting. You should tell me who those people were.

GALISON: There's a book by Timothy Lenoir called *The Strategy of Life: Teleology and Mechanics in Nineteenth-Century German Biology.*

GERSHENFELD: The principle of least action was religious. It was a fight.

GALISON: At the time of Maupertuis, yes.

GERSHENFELD: It wasn't just alternate schools. It was a real religious battle.

SETH LLOYD: Alison, what are these results that you're talking about, about showing that these systems can't supply all the principles? Are these like Arrow's impossibility theorems for voting?

GOPNIK: They've got a very similar structure. Cynthia Dwork is one of the people who has done a lot of work on this, particularly along the lines of thinking about inequalities and fairness. Do you want fairness between groups? Do you want fairness among individuals? They have the same kind of structure as the Arrow theorems, where you literally can't maximize all of those ends at the same time.

To echo what Neil was saying, that's a general principle. We tend to have an idealist picture about computation. It's important to recognize that you're dealing with tradeoffs all the time. That's a very different picture, maybe more like a picture that comes from some enlightenment traditions about philosophy than other enlightenment traditions about philosophy, for example.

WOLFRAM: It's a sad fact about axiomatization of almost anything that people start feeding in all these axioms that they say, "It better be true that this happens and this happens." In quantum field theory, for example, there were these axioms, and then it turned out the only quantum field theory consistent with all these axioms was a free quantum field theory. In other words, there are no interactions between the particles.

TOM GRIFFITHS: There's a sense here in which we're trying to hold machines to a higher standard than we hold ourselves to. Right? This distinction between purpose and mechanism is interesting because we like to think that other people have purposes, but in fact other people mostly have mechanisms. The part of our intuition about moral psychology that's leading

us into problems here is thinking that there is a system that we should be able to formalize and behave in accord with, when in fact none of us do so.

WOLFRAM: There's a little thought experiment that you might find amusing. How does computation relate to democracy? In current democracy, people just say it's a multiple-choice thing. You vote for A, B, C or whatever. But imagine a time when people can routinely speak in computational language as well as in human language, and where it's perfectly possible for somebody to say, "This is what I want to have be the case in the world. I'm going to write this computational essay that is my representation of what I want to be the case in the world." And then imagine that 100 million people take their computational essays and feed them into this big AI that's going to figure out what policy should be followed. That's an alternative to the current version of picking from a small number of choices.

It throws you directly into all of the standard issues of political philosophy of what you are trying to achieve. It's a somewhat realistic view of what could happen, because by the time you have a computational language that can talk about things in the real world, it's perfectly possible for people to represent their preferences in that much richer way.

DAVID CHALMERS: Right here is where you're going to come up against some of these theorems in social choice theory. If everyone's just offering a global vision of the world and we pick one, that's totally unworkable. We've got to find some kind of compromise or consideration of components.

So, we break it down into ten separate issues—A, B, C, D and so on. When we come up against these results and see there's a majority that prefers A and there's a majority that prefers if A then B, but there's not a majority that prefers B. You can't just go with democracy on every component, and then suddenly need some system for somehow extrapolating from all these individual preferences. This is precisely where you need to find ways to make the tradeoffs.

This whole thing of turning morality into code is not a new problem, right? The legal code and the political code has precisely been trying to formalize this for centuries, and what do we know? The only way to do it is via a huge

mess. So, I predict that once you try and turn it into AI code, it's going to be a mess as well.

WOLFRAM: I agree. The main conclusion is that it has to be a huge mess.

WILCZEK: Arrow's theorem ends up with the positive result, which is that the only way to enforce a consistent code is to have a dictator.

LLOYD: That is very positive indeed, Frank. Thank God. Dodged a bullet there.

WILCZEK: The point is that you shouldn't always try to be too rational. Chomsky had this concept, that I find quite beautiful, of crackpot rationalism. Where rationalism is taking you into things that obviously are bad, you should just back off and let the world do its thing.

FREEMAN DYSON
The Brain Is Full of Maps

I was talking about maps and feelings, and whether the brain is analog or digital. I'll give you a little bit of what I wrote:

Brains use maps to process information. Information from the retina goes to several areas of the brain where the picture seen by the eye is converted into maps of various kinds. Information from sensory nerves in the skin goes to areas where the information is converted into maps of the body. The brain is full of maps. And a big part of the activity is transferring information from one map to another.

As we know from our own use of maps, mapping from one picture to another can be done either by digital or by analog processing. Because digital cameras are now cheap and film cameras are old fashioned and rapidly becoming obsolete, many people assume that the process of mapping in the brain must be digital. But the brain has been evolving over millions of years and does not follow our ephemeral fashions. A map is in its essence an analog device, using a picture to represent another picture. The imaging in the brain must be done by direct comparison of pictures rather than by translations of pictures into digital form.

FREEMAN DYSON, emeritus professor of physics at the Institute for Advanced Study in Princeton, has worked on nuclear reactors, solid-state physics, ferromagnetism, astrophysics, and biology, looking for problems where elegant mathematics could be usefully applied. His books include *Disturbing the Universe*, *Weapons and Hope*, *Infinite in All Directions*, and *Maker of Patterns*.

* * * *

FREEMAN DYSON: I was talking about maps and feelings, and whether the brain is analog or digital. I'll give you a little bit of what I wrote:

Brains use maps to process information. Information from the retina goes to several areas of the brain where the picture seen by the eye is converted

into maps of various kinds. Information from sensory nerves in the skin goes to areas where the information is converted into maps of the body. The brain is full of maps. And a big part of the activity is transferring information from one map to another.

As we know from our own use of maps, mapping from one picture to another can be done either by digital or by analog processing. Because digital cameras are now cheap and film cameras are old fashioned and rapidly becoming obsolete, many people assume that the process of mapping in the brain must be digital. But the brain has been evolving over millions of years and does not follow our ephemeral fashions. A map is in its essence an analog device, using a picture to represent another picture. The imaging in the brain must be done by direct comparison of pictures rather than by translations of pictures into digital form.

Introspection tells us our brains are spectacularly quick, transforming two tasks essential to our survival: recognition of images in space, and recognition of patterns of sound in time. We recognize a human face or a snake in the grass in a fraction of a second. We recognize the sound of a voice or of a footstep equally fast. The process of recognition requires the comparison of a perceived image with an enormous database of remembered images. How this is done, in a quarter of a second without any conscious effort, we have no idea. It seems likely that scanning of images in associative memory is done by direct comparison of analog data rather than by digitization.

The quality of a poem such as Homer's *Odyssey* or Eliot's *Wasteland* is like the quality of a human personality. A large part of our brain is concerned with social interactions, getting to know other people, learning how to live in social groups. The observed correlation between size of brain and size of social groups in primates makes it likely that our brains evolved primarily to deal with social problems. Our ability to see others as analogs of ourselves is basic to our existence as social animals.

I go on to talk about what Danny Hillis told us thirty years ago in his paper titled "Intelligence as an Emergent Behavior; or, the Songs of Eden," which is of course a wonderful story that Danny invented to explain the evolution

of speech from song. He had the idea that songs originally were the evolving species, and apes were just the phenotype.

How do songs actually evolve? They have to be remembered by an ape to survive. And how do you get remembered by an ape? Well, you have to give yourself some associated practical use. They have to be useful to the apes in order to survive. So, a song can only become fit to survive by associating itself with meaning. Thereby, you have a co-evolution of apes and songs so that the songs gradually acquire more meaning and the apes acquire more communication. In the end, that develops into speech. This is a beautiful idea. The song is of course analog from beginning to end. It is the sound and spirit of the thing that is transmitted, not the individual phonemes.

I'm suggesting that the brain is mainly an analog device with certain small regions specialized for digital processes. It's certainly not true, as is sometimes claimed by pundits talking on television, that the left hemisphere is digital, and the right hemisphere is analog. It seems to be true that most of the digital processing is done on the left side. But the division of labor between the two hemispheres is still largely unexplored.

* * * *

SETH LLOYD: One of the interesting features in going back over the original Macy Conferences on Cybernetics is that it's a wonderful example of something that is now recurring. The problems that showed up then were somewhat irrelevant for decades, largely because of what Rodney was saying, which is that we adopted von Neumann architecture computers and then Moore's law took off, so we didn't have to bother with different ways of processing information.

They were very concerned about the question of gestalt. What does it mean? Why do human beings get a gestalt—a sense of a whole—from all these disconnected parts? They were questioning what's going on in the brain that gives you this notion of "Aha, that's Freeman right there. I recognize him." They also ask the question of whether artificial intelligences and computers could have a gestalt.

Now, ever since the famous example of Google's deep neural networks learning to recognize kittens on the Internet, at least they have a gestalt of a kitten. Mind you, from a Bayesian perspective, the prior probability of a picture on the Internet being a kitten is rather high. For the first time, it's pretty fair to say that we have artificial neural networks that possess a gestalt. This is amazing, because it's been seventy years since this question first came up. Up until now, I would have said that image recognition programs didn't have the sense of "Aha, it's a kitten," but now they do. So, it's a remarkable time.

FREEMAN DYSON: That's all true. What they call deep learning is imitating this comparison of images by translating to digital language. But still it's not likely that the brain is doing it that way.

STEPHEN WOLFRAM: Neural nets, in their current instantiation, critically depend on the fact that they have real number weights that can be progressively improved by calculus-like methods. It's still an open question as to whether there's a way to do this with purely digital things where there isn't this calculus-like progressive improvement.

F. DYSON: Yes. Certainly, it's an open question. I'm just prejudiced.

WOLFRAM: In your sense, is a neural net with real number weights analog or is that digital?

F. DYSON: That's digital. It's a crude digital imitation of a natural process which was analog.

WOLFRAM: So, to make it analog you would have to have a whole field and not just a matrix of weights?

F. DYSON: Images will slide over each other somehow and match. It's a much more error-tolerant system, so you're not asking for twelve-digit accuracy. If an image looks like another image, then it's essentially remembered together with it. Associative memory is the basis of the whole process, and that works with amazing smoothness that we don't understand.

W. DANIEL HILLIS: Certainly, at some level, there are non-firing neurons in the retina, which are clearly doing a purely analog computation in every sense of the word. If you have something like a Hopfield network, which is basically finding eigenvalues of the matrix by repeatedly feeding itself back into itself, is an eigenvalue a digital output of a completely analog system? Would you put that in the analog category?

F. DYSON: Well, of course you don't have to put things into categories. Most things are a mixture, and that's a good example.

CAROLINE JONES: One of the things that confuses the conversation for me, as an image theorist and a gestalt historian, is that we've made the machines interpolate and extrapolate from the digital to produce gestalt interfaces for us. It's a complicated conversation, because all of the compression algorithms are tinkered to produce something that we will then complete. We will then take the fragmentary pieces and do our analog business on them to create a song and say, "Oh, it's so real!"

We are the cybernetic completion of the digital. We are the analog meat machines that make the gestalt out of what I would imagine the machine doesn't care is a kitten or not. And when you look at some of what Google calls kittens, it's really breaking the gestalt picture. It's a couple of eyes in a certain position and some fur, where the whole premise of gestalt is the completion of the fragmentary, and the curious project by which three different corners are perceived as a triangle, obscured by a circle. The three triangles are robustly perceived as a geometric figure by the human brain, which a machine would only do if we said, "Can you please make these fragmentary corners into a triangle for the human perceiver? Could you please interpolate those missing pieces? We need to see a triangle." So, this interface is productively confused by what we have given the machines as purposes. We have made them into makers of analog maps for us, but I don't yet have a sense of what the machines would do by themselves, for themselves.

GEORGE DYSON: When the Cybernetics Group first formed, that wasn't the name. It was called the Teleological Society. Then, when Macy came in and supported it, he said, "We'll support this, but we've got to have a different name." And that's when they made themselves the Cybernetics Group.

Originally, it was the Teleological Society—that was the fundamental premise.

JOHN BROCKMAN: What would they call this group?

JONES: The Anti-teleological Society.

LLOYD: The Eschatology Society.

FRANK WILCZEK: Post-logical society.

WOLFRAM: What would be the type of theory you would have for what might be going on in the brain? You say it's transforming an image into some different projection of an image, so what's the theory?

F. DYSON: Why did we evolve people like Beethoven and Mozart or Sophocles or Eliot, people who were masters of music or masters of language? This degree of sophistication both in music and in language is far beyond anything that biological survival needed, but it just happened. How do you understand that?

WOLFRAM: You take some simple program, you run it, it does amazingly complicated things, and the program might have been in some sense constructed only because it makes an array of three black cells after four steps or something. It just so happens that as a side effect it produces this amazingly complicated behavior. That would be my metaphor for what's going on in those cases.

F. DYSON: Some quality in the whole scene—the quality of the sunset in the tropics or the quality of a symphony—is just the gestalt, it is something that's inherent in the entire picture and not in the individual parts. That is the brain operating directly on the image and not on the constituent parts.

W. DANIEL HILLIS: The literal answer to your question may be runaway sexual selection. Basically, the way to get laid was to write a sonnet or sing a beautiful song.

ALISON GOPNIK: That may reflect some prejudices in this group. It's not obvious that, generally, artistic and scientific achievement has that effect.

HILLIS: The question is, why are we evolved to support artistic and scientific achievement?

GOPNIK: Here's an interesting possibility, which is something that has come out of the deep-learning world: A lot of times the way you can make those systems work is by having hallucinations, where the system is generating a lot of possible outputs from some representation that aren't actually things that you perceive or aren't inputs into the system.

Having this process of taking a generative model and then simulating a lot of outcomes that you aren't seeing or detecting is a crucial step in making things work. Then, you have another system that looks at the relationship between the generative model and the outputs, and then uses that relationship to the hallucinated outputs—to the things that never existed except that you generated them—and tries to make sense out of that. That turns out to be important computationally.

It's at least interestingly analogous to things like pretend play with children, for example. You don't need to have Einsteins and Beethovens to have examples of people creating things that are non-real. What's the evolutionary advantage to having an imaginary friend or a crazy pretend world? That's not something that you need to depend on experts for. That's something that seems to be a universal characteristic of childhood.

HILLIS: The notion that sexual selection causes you to explore the most complex expressions of those to demonstrate that the complexity is working plays out not just with intelligence but also with morphogenesis. There are all kinds of examples in low level animal behavior, or forms of flowers, things like that, where that process of feedback on sexual selection tends to select for complexity and beauty because that's hard to do. Therefore, it shows it's all working.

LLOYD: If Chomsky were here, he would say that human beings have universal human language, which we gifted to computers, by the way. We're the only entities on the planet that have this universal language. If you look

at chimpanzees, or songbirds, or dolphins, they just cannot process information the way that we do.

One of the features of universal human language is its open-endedness that allows you the potential to construct any possible sets of ideas, or to compute anything in the case of computers. The sonnets and Mozart symphonies, once you give people that, that's what you've got to expect to happen sooner or later.

JONES: I have a different observation, which is that culture is a very unique human product. I'm sure you can argue that bowerbirds have culture and so on, so let's just put that to the side. We have produced these externalities partly to evolve ourselves. That's part of the magic, that you make this thing called art and you gather people around it to interpret it, then they make a certain meaning which then changes them for the future, changes their offspring, changes their survivability rate.

This is part of the operation that fascinates me. Not everybody who listens to Beethoven goes off to have sex with Beethoven. So, what else is going on with art? It is there to evolve us in directions that we agree socially and culturally that we want to evolve. That's rather extraordinary.

NEIL GERSHENFELD: There was an interesting study a couple of years ago that showed birds have hemlines, that they have fashion. What color feathers they have and how long they are changes. There are fashions for the birds. The study traced it through to show that if you didn't do that they would over specialize. If something was considered a locked-in fashion, the birds would exaggerate it. So, they keep having a new hemline to force themselves to diversify.

HILLIS: Part of the appeal of my "Songs of Eden" story that Freeman told is that the "we" that we're talking about is not just the monkeys. The "we" is in fact that culture that evolved. So, what makes us human is that combination of those two things together. What was evolving is not just the genetics that was evolving the monkey, it's the cultural complexity in which all those things should happen, and that's part of what we are. We're the combination of those two things.

IAN MCEWAN: It would seem that all art and all music is a special case of what everyone is doing, so there might be a random element that there are just people who happen to do it better.

F. DYSON: Just one more remark. If you bring in quantum mechanics—of course both digital and analog computers may be classical or they may be quantum—it makes an additional strong advantage to the analog way of working. Quantum mechanics has this quality of coherence that connects parts of the whole physical landscape in this mysterious way; the different parts of an image are coherent. That is totally lost when you digitize, but it's preserved when you do analog. That's an additional reason why analog computing probably looks more promising.

GERSHENFELD: Seth and I were both part of a very interesting program on quantum biology. Biology uses quantum coherence exquisitely, but only over a very small number of degrees of freedom. It's very expensive to preserve coherence. It's very unlikely, and I think Seth you would agree, that there's large-scale quantum coherence anywhere near biology. It's in very selected, small numbers of interacting degrees of freedom.

F. DYSON: No, I disagree totally with that. Quantum coherence works beautifully over large distances.

GERSHENFELD: Over large distances, but it's the question of degrees of freedom and thermalization.

WOLFRAM: What are the examples in biology?

LLOYD: If you just look out the window, all these green leaves are LHC2, which is the primary photo system for plants. It uses quantum coherence in a very sophisticated fashion to increase the efficiency of exotonic transport, and it's amazing. It would be one tenth as efficient if it weren't for this quantum coherence.

GERSHENFELD: Another interesting one is sensing magnetic fields. There's independent chemistry in how you perceive magnetic fields. Maintaining quantum coherence with lots of degrees of freedom against a heat bath is

really hard. That's the challenge in quantum computing. The physics makes it very unlikely there's large-scale quantum coherence.

LLOYD: Well, that's not entirely true. If you look at light from a distant star and you have a big enough telescope, then you can exhibit coherence in this light. This is the Hanbury Brown-Twiss effect, which is what allows you build large baseline telescopes. But that's a situation in which light has traveled, and it could have traveled for millions of years.

GERSHENFELD: And there's no interaction. There's lateral coherence.

LLOYD: It's because it didn't get de-cohered along the way.

GERSHENFELD: There isn't longitudinal coherence, in which case it's lateral coherence.

LLOYD: It's still quantum.

DAVID CHALMERS: Freeman, I'm curious about how you get your model of the analog and the gestalt going without quantum computation. If we assume it's all classical physics and classical computation, then presumably it breaks down into local mechanistic parts.

If I operate on an image via classical mechanisms, it's presumably going to have to work at some level operating on the parts of the image. Aren't you going to come back and say, "Well, that's not what I needed. I needed something holistic that operated on the whole image at once."? One could at least smell a way of trying to do that with quantum mechanics, but how could one possibly do that without quantum mechanics?

F. DYSON: Well, it's just one of the big mysteries. We have no idea how all that works.

CHALMERS: If the brain does it by local mechanisms of neurons, would that count? Or would that still be breaking it down into parts?

F. DYSON: I don't know what a neuron is and neither does anybody else. A neuron is a very, very clever device.

CAROLINE A. JONES
Questioning the Cranial Paradigm

Part of the definition of intelligence is always this representation model. . . . I'm pushing this idea of distribution—homeostatic surfing on worldly engagements that the body is always not only a part of but enabled by and symbiotic on. Also, the idea of adaptation as not necessarily defined by the consciousness that we like to fetishize. Are there other forms of consciousness? Here's where the gut-brain axis comes in. Are there forms that we describe as visceral gut feelings that are a form of human consciousness that we're getting through this immune brain?

CAROLINE A. JONES is a professor of art history in the Department of Architecture at MIT and author, most recently, of *The Global Work of Art.*

* * * *

CAROLINE JONES: I want us to think about the gut-brain axis and the powerful analog system of our immune brain, also thought of as a mobile brain. The cranial paradigm is what I'm here to question and offer you questions about. Mainframe is a kind of discourse that haunts the field that we're talking about, and the cranium comes with that metaphor that we all live by.
What do we mean when we say the word "intelligence"? The immune system is the fascinating, distributed, mobile, circulating system that learns and teaches at the level of the cell. It has memory, some of which lasts our entire life, some of which has to be refreshed every twenty years, every twelve years, a booster shot every six years. This is a very fascinating component of our body's intelligence that, as far as we know, is not conscious, but even that has to be questioned and studied.

As you go to lunch, you will be putting things in your mouth that are not yourself. Your body, hopefully, at this point in its existence, knows better than to reject these not-self proteins and not-self photosynthesizing cells and pitch you into an immunohistological response, and to say, "Oh, this is friend, this should be tolerated. I, this aggregated entity of self, will learn

that these things are friend. These things are to be tolerated, these things are to be learned from and incorporated and not rejected."

Yet, if that same food were somehow injected into your lungs, you might have a violent asthmatic response. You might die from that. The immune system is using the mouth as a category to learn and to train. It took scientists a long time to figure out where this learning and training was happening, which it seems is in the lymph system. If things are introduced there by the injection at the doctor's office, a whole different set of learning is instigated that reads it as not-self, not friend, something which needs to be expunged and eaten by the macrophasias and remembered as not-self, as enemy.

This is an extraordinarily powerful metaphor, and it's one that is parallel with AI and computer sciences now in active transformation. In other words, most of our pharmacological economies are organized around antibodies. But the probiotic industry, which is completely unregulated by the US Food and Drug Administration, is expanding through folk medicine.

When my own immune system was quasi-destroyed and rebooted by chemotherapy, I was like okay, how do I rebuild this? What are the probiotics? What's out there? It's in folk medicine and in proprietary corporate formula. I can know that it's some kind of yeast, but I can't know what the exact sub-species is, owned by this or that corporation, that I'm trying to reeducate my immune system with.

This is a moment of paradigm shift in multiple fields. I'm recommending that we think about the way Catherine Bateson describes some of her father's work, that mind does not necessarily stop at the skin. We are completely symbiotic on these planetary systems that form and have formed our consciousness and our capacities to learn, and to navigate, and to remember. Through our lifetimes, we become hosts, dependent on xenobacteria that we invite into our bodies and cultivate and grow as part of a self that is not yet ourselves, that is a not-self that we cohabit with and are completely dependent upon.

I just throw this out as a complete provocation, which I'm supported by through the cultural evolutionists we call artists, who are making art forms

out of biological materials, out of living materials, to help us think through our symbiotic dependence on other life forms and our interesting non-conscious negotiation with self and not-self every day. I can be very brief and leave it at that.

Frank's comment about just being in the world, embedded in an environment, and letting that surfing and negotiating with inputs that are analog and need to be responded to in an adaptive and flexible way—this is what I would call intelligence. The body is an amazing model that goes way beyond mind of learning and memory and how we can craft our epigenetics through certain cultural acts and practices, how we can supplement them prosthetically, epigenetically.

This is my provocation to reboot AI on a certain model of what Gibson would have called "environmental and ecological perspective."

* * * *

STEPHEN WOLFRAM: I'm curious as to what the current immunologists' high-level model of the immune system ends up being. For a long time there were these network models of the immune system where there are antibodies and anti-antibodies, and there was this notion of dynamic equilibrium in the immune system. If you ask a random immunologist what their high-level view of the immune system is, in my experience, they'll tell you a very low-level view of a very specific part of the immune system. There used to be these network theories of the immune system. It would be interesting perhaps to compare those with the current models for brains and neural nets. I'm just curious if people know what the current view is.

JONES: I'm not going to be able to answer that. Are you speaking of theoretical biology? Are you speaking about practical immunologists in a hospital setting?

WOLFRAM: There are 100 billion possible types of antibodies, and any particular person has some number of those antibodies in reasonable concentrations. You can do an assay for a particular antibody, but this question of how many of the 100 billion possibilities do we have in decent

numbers, I don’t think that’s known. You start getting more and more antibodies, why does it not run away?

JONES: The obsession has been on the antibodies, and that is part of the systemic immune system that has received all the research. What hasn’t been researched is the mucosal immune system, which is the system that learns, the system that builds tolerance, the system that trains and takes in and negotiates the self/not-self.

Part of the example I'm about to share comes from a neuroscience boot camp I took at Penn, which was great. In the presentation somebody said, "Oh, then there are the glia." And I said, "What are the glia?" And the response was, "They’re not important. They’re the house cleaning staff." I said, "I’m a feminist. The house cleaning staff is important to me. And by the way, your model of mental illness is a serotonin reuptake inhibition model, so you’re dependent on the cleaning crew to manage this uptake."

Basically, there was this tiny window into an under-researched entity in the brain that is entirely involved in the immune system. It used to be thought that the brain is somehow isolated in its beautiful ivory cranium, and it just doesn’t have to deal with the immune system, and it’s kept away from all those diseases. Well, no, the glia are there actively cleaning up, managing the garbage that is produced by the phasias that are eating the toxins and determining what is self and so on and so forth. I believe, not being a scientist in this world, that they are at the edge of shifting into some very different kinds of research not on the heroic actors with their shields and swords, but the clean-up crew that is determining how the body will go forward.

RODNEY BROOKS: You talked about the immune system as a separate system. We have the gut neurons, which are separate. Even C. Elegans (Nematode), which has 302 neurons, twenty of which are in a separate gut-brain than the central brain, and they have fifty-six glia cells. Even in that smallest thing, we see this structure. You were talking about the immune system in an interesting way, as learning, teaching, and remembering. When we look at plants and their capabilities, they don’t have neurons either, but the roots go out and search and the leaves do all sorts of things. There's a lot of activity that is underappreciated when compared to the neurons, which

are seen as the only important cells in the brain. It's even worse when you think about non-animals, because there's stuff happening for which you don't have computational models, which gets back to my earlier point. When do we have computation models and when don't we? Plants are obviously doing something very interesting in the way they adapt to their local environment and adapt to what's happening and change themselves.

ALISON GOPNIK: This gets back to something that you were talking about, Rod, regarding adaptation. There's one dimension, which is one of the things that Turing realized, and that's the idea of breaking up something complex into a process where you can describe it as parts of the process. That's the big idea of computation. One kind of intelligence is being able to do that.

Another idea is being able to represent something that's external to you, being able to take the external structure of the world and, in some way that we don't quite understand, get a veridical account of what's going on in the world around you or adapt to what's going on in the world around you. There are interesting questions about what the relationship is between those two kinds of intelligence. And they might be orthogonal to one another in various ways.

Something like deep learning solves this problem of trying to adapt to the external world in a very simple way. It lets you take the statistical structure of the input, something like images on the Web, and incorporate those into a system that's producing a particular kind of process. But that's a primitive way of relating to the external world. I don't think we have a very good theoretical account of how that process of adapting to the external world is related to the process of being able to compute. I don't think we have a good story.

JONES: That's true. In parallel to the neuron supremacists, we would have the representation fetishists. Part of the definition of intelligence is always this representation model. When we think about the immune brain, I don't think we need to imagine the glia having a representation of the body or even a map of the body. It has pathways to circulate in. It might even respond to some vessel making components of the body to make new pathways, if it needs them. I don't know those mechanisms.

The point is, it does not need a representation of the body; it needs to know where it needs to go, which is a different problem. I'm pushing this idea of distribution—homeostatic surfing on worldly engagements that the body is always not only a part of but enabled by and symbiotic on. Also, the idea of adaptation as not necessarily defined by the consciousness that we like to fetishize. Are there other forms of consciousness? Here's where the gut-brain axis comes in. Are there forms that we describe as visceral gut feelings that are a form of human consciousness that we're getting through this immune brain?

PETER GALISON: When the bio-artists look at microbiological forms and plant forms or animal forms, is there something suggested among them that might give us a different set of metaphors or conceptualizations of consciousness?

JONES: Through the artists, I'm coming up with this idea of symbiontics, "ontics" being that which is, and symbiosis being that which I wish we could be more completely aware of as we navigate this world. Many of them work with concepts and materials in the gallery that prompt me to think more robustly about our interdependencies.

GALISON: So, what are the artists doing?

JONES: Philippe Parreno, who we saw in Berlin, used bacterial motors to turn the lights of the gallery on and off and raise the window blinds. The bacterial motors are entrained with other forms of AI and digital computations that are responding to the presence and absence of humans and their movements through the space, as if we were invaders of a non-self that the gallery must then respond to as an immunological distributed system.

You could think of these metaphors in lots of different ways, but the artists are helping us evolve toward a more symbiotic understanding of our place.

DAVID CHALMERS: One angle on thinking about self and not-self in cognition and intelligence? We could bring in the literal immune system, as you've done, but we could also think about it in terms of having a separate cognitive immune system, drawing the self/non-self distinction at the

cognitive level. A lot of that is done by things like trust. If you ask about self and non-self in cognition, my smartphone is totally self. It's not non-self. It's not something outside which is coming in.

JONES: You never get spam calls?

CHALMERS: It's app sensitive or context sensitive. The phone numbers and Google maps, that's just self, that's my navigation system. I treat it as self and I trust it. It basically becomes an extension of my cognition.

JONES: So, it's a prosthetic self.

CHALMERS: Yes, it becomes prosthetically part of the mind because I choose to trust it and identify it as self. And there's other stuff out there. Spam, for example, that comes in over email, and who knows what on the Web—these things that I regard as not-self are no longer part of my cognition. This is the way that cognition gets distributed out from our brain into the environment.

JONES: Part of what I'm advocating, and what the art is helping me to do, is advocate for a much broader self. In other words, we know that humans have evolved clothing, and language, and heating, and H-vac, and architecture; if we take these things away, we'd survive, I don't know, maybe two weeks.

In other words, it's partly to acknowledge our existence as social animals, to acknowledge that the cranium is not where we do most of our thinking and being, and to figure out how to get our artificial systems, our prosthetic systems, to help us acknowledge our embeddedness. Partly I see this as a planetary dilemma. If we don't feel our place in the planetary ecosystems, we deserve to go extinct, which we will.

ROBERT AXELROD: I was going to ask whether you think hormones provide another form of intelligence, adrenaline for example.

JONES: Oh, absolutely. Psychiatric diagnoses are being made on the basis of which drugs you respond to, which are influencing hormone cycles and their reuptake by the brain. If we look at how we're practicing this medicine

and dealing with this thing we call the brain, then the mind is totally distributed throughout the body. Hormones are very much a part of that.

IAN MCEWAN: Can you say something about the civil war that occurs when the body turns against itself—Crohn's disease, arthritis?

JONES: There's an argument for ingesting in the oral tolerance portal that which our immune system is turning against. For example, collagen, in the form of certain autoimmune diseases. It doesn't have to be human collagen. There's enough molecular similarity between cows and chickens that if you ingest, rather than inject, this form of collagen, your body is like, "Oh, I don't need to attack that. That's an okay thing, collagen."

Why is the body attacking its own collagen in the first place? If mice are given neural sheets without adjuvants, without things that are alarming their immune system, they will stop having MS or something, their sclerota will stop being attacked by their body. So, there are incredibly promising therapies that are emerging from this.

MCEWAN: But the system can make mistakes.

JONES: The system absolutely can make mistakes. If you think about the form of the vaccine, you think you're just getting polio, but you're getting polio surrounded by a witch's brew of cholera and diseased bacteria, and things that are saying to your body, "This is really bad! Turn against this!" It's the adjuvant. And what adjuvants are we taking in? Let's hope Monsanto isn't on those fields out there. The pollution, the pesticides—these are adjuvants that are alerting us to attack certain things as toxic that then may be disrupting other parts of our immune system. Again, this is still in the realm mostly of folk medicine.

There's an idea that you can reduce asthma if you eat local honey, because what you're eating is the bee's concentrate of all the airborne proteins and pollens and dust, so you're eating all of your local airborne potential triggers. You are learning and training your body to tolerate them by eating the local honey.

Urban asthma is very bad, so should you be eating cockroach feces, or should you just get rid of the cockroach feces? This then becomes a social problem. I'm just recommending this as a model of intelligence which is quite distributed, not conscious. Does it have ethics? It certainly has goals, but they shift every day depending on what part of the system encounters what not-self element.

AXELROD: I would say the immune system has pretty stable goals, which is evolutionarily to protect the host.

JONES: That sounds like a reasonable thing.

AXELROD: The methods change every day, depending on the challenge, but the goal is pretty stable.

JONES: We assume cancer was always there, so why was that not evolutionarily eliminated? We're looking at a system that didn't completely eliminate these things that seem like they would have been evolutionarily problematic. So, is the goal wrong or is the system messed up? Or do we just need to see this homeostatic navigating as part of life?

AXELROD: Are you saying evolution is not finished?

JONES: Well, let's hope it's not. Do we want evolution to be finished?

WOLFRAM: Are you suggesting that the reason there's a rise in autoimmune diseases is because there are more adjuvant-like things in the environment?

JONES: Well, that is the suggestion on the table. It's not my suggestion.

WOLFRAM: What's the leading suggestion for what the adjuvants are?

JONES: What we've described as pollutants is probably a pretty good category: Benzenes, BTEX chemicals, pesticides, things that we have produced either as by-products of energy or to kill lifeforms. Rachel Carson wanted to call these biocides, because they're not just killing pests.

WOLFRAM: Epidemiologically, that would be a fairly easy question to test I would think. That's an interesting theory.

JONES: They're now beginning to test this. You have to understand that you're dealing with industrial food and industrial medicine where those haven't been the leading research questions. Monsanto doesn't want us to ask about that. Monsanto wants us to buy Roundup, so it's hard to get that research done into whether people living in blow fields near Monsanto are having more autoimmune diseases than people who don't. It's hard to find people who are not living near Monsanto drift. These are important questions, and they are starting to be tested.

ROBERT AXELROD
Collaboration and the Evolution of Disciplines

The questions that I've been interested in more recently are about collaboration and what can make it succeed, also about the evolution of disciplines themselves. The part of collaboration that is well understood is that if a team has a diversity of tools and backgrounds available to them—they come from different cultures, they come from different knowledge sets—then that allows them to search a space and come up with solutions more effectively. Diversity is very good for teamwork, but the problem is that there are clearly barriers to people from diverse backgrounds working together. That part of it is not well understood. The way people usually talk about it is that they have to learn each other's language and each other's terminology. So, if you talk to somebody from a different field, they're likely to use a different word for the same concept.

ROBERT AXELROD, Walgreen Professor for the Study of Human Understanding at the University of Michigan, is best known for his interdisciplinary work on the evolution of cooperation. He is author of *The Evolution of Cooperation.*

* * * *

ROBERT AXELROD: Let me start with what's new in the world of cooperation. There's the problem of international relations in which an established power, the United States, is dealing with a rising power, China. The ancient Greek historian Thucydides said that the reason why Athens and Sparta fought was because Athens was a rising power and Sparta was the established power and they couldn't work it out. More recently, Graham Allison at Harvard looked at the last 500 years for all the cases in which an established power was dealing with a rising power. He found sixteen of them, twelve of which led to war. Those are not good odds.

One of the ways of dealing with this is to try to develop norms and rules of the road for understanding what's proper behavior. I'm working with Chinese and American delegations who are meeting regularly to discuss things like

cyber conflict. For example, if cyber weapons were used on a large scale, it looks unstable in a way that nuclear weapons are not unstable. So, we're dealing with how to develop norms for understanding cyber tools and cyber weapons. That's one area where cooperation is important.

Another area, which you're all very familiar with, is the decline of democratic norms not only in the United States but in many other countries, especially in Europe, where the basis for societal cooperation in a sense of governance are deteriorating. A third area is climate change, where one could certainly look at this as a technical problem. I hope technical progress can be made, but it's also a collective action problem of getting large numbers of actors to work together.

Interdisciplinarity is another area where cooperation is needed and is not trivial to attain. The research on what makes interdisciplinarity succeed when it does and what its characteristics are has exploded in the last ten years, in part because of the ability to do large-scale analysis of things like citations and see whether people who publish articles together from different disciplines are more successful in, say, achieving citations.

There are a few things that are known. One is that interdisciplinary research has higher variance. It's not a higher average of success, but it's higher variance, so sometimes it does very well and many times it does not so well. So, it's not necessarily that more interdisciplinary work is better. If we could understand better barriers to make it work, we could maybe change that.

Another finding is about preferential attachment. The idea is if you work with somebody, you're likely to work with them again, and maybe even second-order, where you'll work with people that worked with them again. Another result is that you can map the disciplines in two dimensions such that distance represents the probability of collaboration; for example, you might have a lot more collaboration between, say, physics and chemistry than you would between physics and sociology. That's not surprising, and the maps look plausible and reasonably stable. But that's not a lot of knowledge about interdisciplinarity.

The questions that I've been interested in more recently are about collaboration and what can make it succeed, also about the evolution of

disciplines themselves. The part of collaboration that is well understood is that if a team has a diversity of tools and backgrounds available to them—they come from different cultures, they come from different knowledge sets—then that allows them to search a space and come up with solutions more effectively. Diversity is very good for teamwork, but the problem is that there are clearly barriers to people from diverse backgrounds working together. That part of it is not well understood. The way people usually talk about it is that they have to learn each other's language and each other's terminology. So, if you talk to somebody from a different field, they're likely to use a different word for the same concept. That also comes up with Americans talking to Chinese about military things. That seems to me just part of it.

Another part of it is whether they have common goals. For example, if there are two different disciplines, the researchers might want to publish in journals from their own discipline so that their own peer group will recognize the contribution, and that could be a conflict of interest between them that they need to work out.

The other problem is that what they come up with in a collaborative interdisciplinary activity may not be recognized as a contribution by any field, and this is especially true when there are new fields. One of the things, though, that does make the interdisciplinary activities that I've been involved with work is having some tools in common. For example, game theory is understood as valuable and taught in much of the social sciences and the biological sciences, so being able to collaborate with someone who knows game theory gives us a chance to make progress.

Ian mentioned civil war in a body as an interesting aspect. I saw an agent-based simulation of a growing cancer where the agents were the cells. I asked the computer scientist that developed that with a student, "What are the premises that go into that simulation? What are the mechanisms?" They pointed me to an article on the hallmarks of cancer, and it turns out that there are about eight different defenses that the human body has to keep cells in line from becoming selfish and asking for more resources than is good for the host. The common understanding was that a single cell line develops the mutations necessary to overcome each of those defenses, but

when I saw the simulation and read about the mechanisms, I realized that that wasn't necessary.

Let me give you an analogy: If you have two thieves robbing a house and one of them knows how to turn off the alarm and the other one knows how to pick the lock, they don't both have to know how to overcome the defenses as long as they're traveling together. In cancer, some of the defenses are overcome by putting out a certain chemical saying, "Build a capillary in my direction," basically asking for more blood and more oxygen, but another cell nearby might be exceeding its normal capacity to do something which would overcome another defense. So, as long as they're together, you don't need a single cell line. That would help give you another channel for therapy, which would be to interrupt the cooperation in the cell line.

I went to a geneticist and an oncologist, and we worked out some of the implications of this. First of all, we found that it hadn't ever been explicitly stated and, secondly, it was biologically plausible, so we wrote this up. So, here was collaboration between a political scientist, an oncologist, and a geneticist. When we proposed this speculation about how things might be, we got two reviews. One of them said what we were proposing was impossible, and the other one said what we were proposing everybody knows. Anybody had a pair of reviews as challenging as that? Obviously, we didn't explain ourselves very well, so we picked ourselves up off the floor, rewrote it, and tried to explain why it's neither impossible nor the same thing that people knew.

What provided the basis for the collaboration is that I was looking at this from a social science perspective of community action and cooperation, and the others had the competence about how cancer works. We were working on a known problem, which is, for example, how does civil war in your body get under control and how does it lose control? One of the opportunities for success is if there's already a known problem and then you provide another way of attacking it or making progress on it. As long as the problem is accepted in at least one discipline, then it seems to me you could use any tools and any new concepts as long as you could make progress in the terms that the people that care about that problem understand. If it's a problem that they don't realize they have, it's much harder.

Let me talk about the evolution of disciplines, which is one way to think about this. One way to approach disciplines is to see them as an ethnic group or a language group where the people within a discipline are able to talk to each other well. This is because the disciplines have become institutionalized so that anybody that calls themselves an economist or geneticist knows a whole bunch of stuff that almost every other economist and every other geneticist would know. So, they can talk to each other in the areas that the discipline has defined as building on their canon. That's fine, but it's like a gravity model in a sense that then the disciplines become more and more coherent over time, and that makes it easier. Then there's a body of concepts, and terminology, and science, and previous experiments that are shared, and that makes that kind of collaboration easier. But there's another group over here that has coalesced in a different place in this high dimensional space. As they each coalesce, they become further apart with fewer people in between.

Another analogy might be like Spanish, French, and German. There used to be a whole series of dialects that are more or less continuous across that space. Eventually, those three countries established the canonical way of saying German, French, and Spanish and taught it in the schools, which was very useful in the Industrial Revolution when you wanted people from a distance to be able to deal with each other. Then it wiped out most of the stuff in between, Catalonia being a surviving exception.

In disciplines we've converged, the convergence is not just on subject matter; it's incredibly well institutionalized so that departments not only represent disciplines like physics and economics, they also control careers. They decide whom to hire and, therefore, professionals have a strong need to be attractive to at least one of those disciplines. Not only that, but they control the entry. They control the training process to determine what it takes to get a PhD in X or Y. They also control, to some extent, the journals and the major professional conferences. They don't control the smaller journals or smaller conferences. So, when a group like us gets together with different backgrounds and tries to communicate, there are several questions about whether there's an emerging discipline of brain intelligence and neuro and cognitive psychology, because all those disciplines are so well established and institutionalized. It's not easy.

It’s easier to get a center going perhaps, but it’s very hard to then get the established groups to give tenure lines, faculty, resources, course credits, and the ability to grant the PhD under their label. The way these things have coalesced to some extent is accidental, though not completely because there is a difference between what chemists study and what physicists study—in the matter of scale, for example. In other fields it’s not as obvious where the boundaries would be if you started over again or, more to the point, where they should be now.

You can’t erase the boundaries and just redraw them. Several places have tried that. Carnegie Mellon and Irvine are famous for having redrawn boundaries, and you can see that the problems they have include the fact that they can’t develop a cascade of reorganization across the academic community. So, they're at a disadvantage, say, on whether their PhDs are hirable and whether the cluster of things that they teach in one of their structures doesn’t correspond to what anybody else does.

The evolution of disciplines seems to take several forms. One is the splitting off of a single discipline into several disciplines or usually one. Maybe astronomy is on the edge of being separated from physics in some places and not others. Clinical psychology is quite different from developmental or cognitive psychology, but they’re still holding together.

Sometimes a new discipline can arise from the territory between. Biochemistry in some places is a new discipline. One of the constraints that helped define this is how much can a PhD candidate learn in five years? They can learn a set of tools, concepts, and experiments. When a single discipline is in the situation where some of it takes five years to learn and other parts take a different set of five years, then it’s pretty ripe for separating those things out and giving them different names and then having fission, and that’s certainly one way that it happens.

Another way, though, is more typical of this room, which is where people from many different disciplines are working on some problem area, like questioning what intelligence is and how the mind works, and how can we accomplish more effective AI and what would it mean to do that. So, we can gather together in this room and try to understand each other, which is certainly a significant task that can be promoted by having repeated

meetings of largely overlapping people, but it's hard to buck the established institutional frameworks.

Caroline mentioned just before lunch the topic of immunology and how the immune system has a kind of intelligence. Let me give you another collaborative example dealing with the immune system. I had worked with evolutionary biologist Bill Hamilton on evolution of cooperation in biological systems. A couple years later, he came to me and he had a theory of the origin of sex, and the theory was that it's an adaptation to resist parasites. That seems very strange, and it goes like this: Parasites have an incentive, a biological selection pressure, to look as much like you as they can. If they look like you, your immune system will not identify them as non-self. You can imagine a high-dimensional space, basically the antigens in which you're located here and the parasites can evolve to become more and more similar to that and eventually get to the point where you don't recognize them as foreign. They have an advantage because they can reproduce perhaps 100 or more times faster than you can. So, they can outrace you as you run away from them. When I mean you, I mean your progeny over generations.

Bill Hamilton's idea of what sex does for you says there's one adult here and one adult here, and they're quite different in how they present themselves to their immune systems and to parasites. If you could take some of the genes from this one and some of the genes from that one, you've made a huge jump in this high-dimensional space. You haven't just moved incrementally. If you had asexual reproduction, your children would be very much like you, but if it's sexual, then from the point of view of the antigen you're very different. Therefore, sex is an adaptation to resist parasites. The problem it has to account for is that only half of the adults have offspring. This is a tremendous biological disadvantage. It could be up to two for one. That's a lot to overcome, so there's got to be some powerful things on the other side to show that at least it's plausible. He said he tried to model this as being explicit about the three characteristics of your adult, and then if you have three others from the other adult and you mix those and then model that, but he couldn't do the math after about three. It doesn't work for three.

I learned about the genetic algorithm from John Holland where you can have long strings of chromosome simulation. "Seventy is no sweat," he said,

"that's just what I need." So we did some simulations, which were enough to make the search problem hard or to make sex valuable.

The problem of why we have sex is a well understood problem. It's well understood that that's a serious problem in Darwinian theory because the two for one disadvantage is so great. There's another explanation for why sex is the answer, but this one looks pretty cool. This allowed me to take something from computer science search techniques and adapt it for a simulation of an evolutionary biology technique.

We are now faced with the question of intelligent AI systems, and that is a lot like disciplinarity. The humans have some set of concepts, and the artificial intelligence system will have another set of tools, concepts, ways of organizing the world, and thinking. How can we promote the effective collaboration of humans and intelligent systems? Then the other question is, how do you guys do it, and what is your experience with effective collaboration across disciplines?

* * * *

STEPHEN WOLFRAM: Has anybody made a giant wall chart of the evolution of disciplines over the last few hundred years?

AXELROD: I've looked for that, too. No. There are some histories of universities that have been around a long time, like the University of Padua, which separated philosophy from law. But I haven't seen it more generic. That would be interesting to do.

NEIL GERSHENFELD: Stephen, could you derive it from all the data you've ingested?

WOLFRAM: I was wondering. The Web of Science, through which you get the Science Citation Index is hard to get access to. There's now this open citation project where journals are contributing their citation metadata, which started maybe two years ago or something and it's gathering steam, but that data is mostly fairly recent data.

GERSHENFELD: Erez Lieberman Google Book data, and he did a surprisingly good job of deriving history.

WOLFRAM: In academia, there isn't management of research. In a company, there is management of research. It's interesting what the tradeoffs are between managed research and unmanaged research.

CAROLINE JONES: Historians like to see the pattern of disciplines as infinite proliferation. So, the phrase "renaissance man" was invented in the 17th and 18thcentury to describe a lost nostalgic moment of wholeness, when you possessed all disciplines in one person. It was already acknowledging that there were divisions happening.

PETER GALISON: What do you see as the main difference?

WOLFRAM: For my company, and I can't say this very scientifically, but over thirty-something years we've developed the concept that we're going to put together these teams of people with different expertise and they will work together. In earlier times, that was hard to achieve, but we finally got to the point where, culturally, we expected that people from different backgrounds would work together in teams. I don't know whether that happens in universities as effectively. It's something that took a long time for us to achieve.

DAVID CHALMERS: I don't remember who it was who said the best interdisciplinary conversations take place inside one person's head.

IAN MCEWAN: I attended a university in which the vice chancellor's project in the early `60s days of great optimism in Great Britain was to redraw the map of learning. Interestingly enough, every student of the humanities was required to read three books. One was Turner's Thesis on the expansion of the American West, one is Jacob Burckhardt's *Civilization of the Renaissance,* and the other was Tawney's *Religion and the Rise of Capitalism.* This was based on the understanding that you could not approach the humanities without a background in historiography, the nature of history, and the changing ways in which history is studied. That whole project lasted about fifteen years and then was swept away. Now, when I go

back to my old university, there's the History Department and the English Department—a kind of inertia dragged it back.

GEORGE DYSON: You already gave the answer to your question, or Hamilton did in his beautiful explanation for the origin of sex, which was the same reason you should have these interdisciplinary things because it allows you to outrun the parasites who build up in the History Department.

SETH LLOYD: Sex is the ultimate interdisciplinary act.

JONES: It speaks to the internal problem of errors in replication that accumulate without hybridization. Self-replication without hybridization risks a lot of errors and repetition of errors and accumulation of errors.

WOLFRAM: You see a discipline in its first generation, the people who founded the discipline are still around; they still know what the fundamental questions are; they're still often a bit insecure about the foundational things that the discipline is based on. Then you get to more generations, and by the time you're at third generation of people, typically, they don't even discuss the foundations. It is just assumed.

If you look at the period of maximum fertility, maximum lasting effect of a discipline, is it the case that most of it is in the first ten years? Is it in the first twenty-five years? Is it in the first generation? Should disciplines be euthanized after they've gone through five generations, for example? They basically won't produce significant output by the time they've gone through five generations of people.

AXELROD: I doubt that. It seems to me that's like saying, should we get rid of some culture because it's been around a long time and it's worked as much out as it can. That seems silly.

WOLFRAM: That's an extreme version, but my question is, at what point in the curve? It's true with conferences, for example.

AXELROD: Paradigms can get mature and stale, but disciplines can change paradigms, and that's a possible form of regrowth. Let me give you an interesting example of biology where the disciplines of biology have been

now reorganized to turn 90 degrees. It used to be botany and zoology—plants and animals. Now it's skin in and skin out. Skin out is ecology and evolution and skin in is microbiology, so they just turned the whole thing sideways. That's a case where there still is more fluidity in the structure of biological disciplines than there are anywhere else. Maybe that's because there's more things being discovered faster.

The disciplines are often tool-based. Microbiology wouldn't be possible without microscopes. As we get tools to deal with artificial intelligence, for example, then it gives us an ability to see psychology in new ways.

TOM GRIFFITHS: An analogous rotation is happening in psychology, where you have neuroscience as one set of methods, which includes social neuroscience and clinical neuroscience, and then behavioral psychology is another set of methods. Those things would have traditionally been clinical psychology, cognitive psychology, social psychology, and so on, but then getting rotated around into behavioral methods versus neuroscientific methods, where the questions that are being asked in those groups are now starting to be quite different.

AXELROD: One problem with a new discipline or new restructuring is the evaluation of performance. People in the new field will tend to say that anybody in the new field is worth valuing, worth promoting, worth funding. And because it's a new discipline, it doesn't have a well-established hierarchy that you can look to and say, "That journal is the best in the field." Even if you could claim that a journal is the best in this new field, you can't necessarily say how it would rank among other things in a broader area.

An ability to do high quality and consensual evaluation outside of a new area means that there's a comparative advantage for somebody not very good to go into the new area. That in turn leads to a suspicion of people in the old area that the new area is attracting people that couldn't successfully do the old thing, not that they weren't interested in it or that they were pioneers in some sense, but maybe because they see they would thrive in some area where the evaluation is harder.

LLOYD: Sometimes what can happen to an old field, particularly those in which the original founders of the field might not have been the nicest

people in the world, is that they can get grumpy and refuse to cite each other. Who was telling us about an NSF panel where there are these three solid-state physicists on this NSF panel, and they recommended none of the proposals be funded.

WOLFRAM: One thing I've noticed in the creation of new fields is that you ask who are the people who come into a new field when it's created? For example, are they young people? Are they old people? What's the type of person who comes into a new field when it's created? I've been surprised that it's not just young people who come into a new field. It seems to be the case that there are fields or paradigms that are suitable for particular individuals for one reason or another, and if they're lucky enough to live in a time in history when their paradigm is one that is being pursued, then they gravitate to that and they start doing it.

My anecdotal observation is that if you look at people who come into a new field when it's young and then you wait twenty years or something, about half the people who came into the new field when it was young are still in that field, and the other half have gone on to do three new fields or something after that. I'm curious what other people's experience is.

ALISON GOPNIK: There's an interesting tension that comes up with what Tom was saying about when fields bifurcate in terms of methodology or when they bifurcate in terms of content. In some ways, the methodological differences make it harder to communicate because the tools are different.

On the other hand, my experience has been that successful interdisciplinary cases are those that occur when you get people using different methods who were trying to answer the same question. Sometimes you get questions that are even narrower than the question you'd typically think of as being a domain question. For example, rather than asking how we're going to solve the problem of the mind, we ask how we're going to figure out how causal inference works. That's a nice example of where we succeeded in getting real interdisciplinary work.

Another example is the question of how are we going to figure out how people understand what's going on in other people's minds, which came to be called theory of mind. That promoted genuine interdisciplinary work, and

that was because you had people with different methods who were trying to solve the same problem. I can't even think of examples where what would happen is that you get a bunch of people together because they said, "We're all using the same methods and we want to find out more about the common methods even though we're solving different problems."

AXELROD: Does anybody have examples of collaboration across disciplines that didn't work?

LLOYD: Like you, I've done a lot of interdisciplinary work and effectively invading a number of fields. Some fields very much don't like being invaded, but also at the same time they're flattered that somebody is paying attention to them. Starting about ten years ago, I and some friends from the quantum information community realized we could make some reasonable contributions to this field of quantum mechanics and photosynthesis. This is a small field full of grumpy old men who never cite each other's work. Science proceeds one death at a time. At one of the conferences I said, " I've never met a field so closed in. You can only make progress in your own specialty by dying yourself," which graduate students thought was very funny, but the professors didn't think it was very funny.

AXELROD: Your problem of how does photosynthesis work was obviously accepted as important. Photosynthesis is obviously accepted as an important thing and how it works was understood. It was already understood that they didn't have a complete satisfactory account and, therefore, if you could provide a better account then they could appreciate that that's a contribution no matter what tools you used to get there.

JONES: But he's saying it was cranky and they didn't accept it.

LLOYD: It ended up being accepted. They did need to learn about it because they didn't understand what was going on with quantum coherence, and the methods that we supplied did allow them to figure that out, but boy they were dragged kicking and screaming to this and still don't want to cite the work.

JOHN BROCKMAN: You mentioned that you've gone back to science rather than writing this book you were talking about. So, what is your science? How

would you define advances in the science? I don't quite understand what the field is.

AXELROD: I guess I'd put it two ways. The core of my interest is in international relations, especially great power relations and issues in times of war and peace. I'm also a math modeler. I'm looking for opportunities to do math modeling, in particular, agent-based modeling and Santa Fe Institute kind of complexity work. I'm an opportunist and a curious person, so if I see a model of a simulation of cancer, I try to figure out how it works. I make a real effort to meet people and talk to them often over lunch and sometimes at meetings like this one.

So, my field could be something different. I haven't done anything specialized in artificial intelligence, but I'm fascinated by it. As a social scientist, I see that one approach to autonomous vehicles—which are our best example currently of sophisticated artificial machines—is to make them more sophisticated and better able to understand the environment and avoid mistakes. A whole other approach involves legal liability questions. The problem is, who's going to take responsibility for the accidents, and how do we institutionalize that judgment between the insurance companies, between the manufacturers, between the owner of the car, between the person that sets the parameters?

LLOYD: Would you recommend to a junior faculty member to pursue interdisciplinary work?

JONES: There are fields that privilege single author and fields that privilege multiple authors, and I think the answer would be different based on these two models.

AXELROD: In order to get tenure, the tenure committee wants to know how good a bet you are in the long run. Let's say all of your work is done with the same senior person. Well, that's not a way to build up a record that then could be evaluated. If you're going to do collaborative work, you should collaborate with different people to so that your work is distinctive. The other is to make sure that you're single-authored work is among your best work. Work with different people and make sure that your single-authored stuff is among your best

ALISON GOPNIK
A Separate Kind of Intelligence

It looks as if there's a general relationship between the very fact of childhood and the fact of intelligence. That might be informative if one of the things that we're trying to do is create artificial intelligences or understand artificial intelligences. In neuroscience, you see this pattern of development where you start out with this very plastic system with lots of local connection, and then you have a tipping point where that turns into a system that has fewer connections but much stronger, more long-distance connections. It isn't just a continuous process of development. So, you start out with a system that's very plastic but not very efficient, and that turns into a system that's very efficient and not very plastic and flexible.

It's interesting that that isn't an architecture that's typically been used in AI. But it's an architecture that biology seems to use over and over again to implement intelligent systems. One of the questions you could ask is, how come? Why would you see this relationship? Why would you see this characteristic neural architecture, especially for highly intelligent species?

ALISON GOPNIK is a developmental psychologist at UC Berkeley. Her books include *The Philosophical Baby* and, most recently, *The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children.*

* * * *

ALISON GOPNIK: Everyone knows that Turing talked about the imitation game as a way of trying to figure out whether a system is intelligent or not, but what people often don't appreciate is that in the very same paper, about three paragraphs after the part that everybody quotes, he said, wait a minute, maybe this is the completely wrong track. In fact, what he said was, "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child?" Then he gives a bunch of examples of how that could be done.

For several years I've been pointing to that quote because everybody stops reading after the first section. I was searching at lunch to make sure that I got the quote right, and I discovered that when you Google this, you now come up with a whole bunch of examples of people saying that this is the thing you should be quoting from Turing. There's a reason for that, which is that the explosion of machine learning as a basis for the new AI has made people appreciate the fact that if you're interested in systems that are going to learn about the external world, the system that we know of that does that better than anything else is a human child.

One of the consequences of that, which is not so obvious, is thinking about children not just as immature forms who learn and grow into an adult intelligence, but as a separate kind of intelligence, which is implicit in the Turing quote. That fits with a lot of interesting ideas in evolutionary biology.

In evolutionary biology there's increasing work on the idea of "life history," but if you talk to developmental psychologists, they've never even heard of it. Life history is the developmental trajectory of a species: how long a childhood it has, how long it lives, how much parental investment there is, how many young it produces. That general feature of what its life history is like is often much more explanatory of other features of the organism than things that might seem to be more apparent; in particular, a relationship that comes up again and again is a relationship between what we perhaps anthropomorphically think of as intelligence, things like being able to deal with many different kinds of environments, learn about them, and adapt to them effectively. That turns out to be very consistently related with a particular life history pattern, namely a life history in which there are few young, a very long period of immaturity and dependence, and a great deal of parental investment.

The strategy of producing just a few younger organisms, giving them a long period where they're incapable of taking care of themselves, and then having a lot of resources dedicated to keeping them alive turns out to be a strategy that over and over again is associated with higher levels of intelligence. And that's not just true for primates. You can see this in analyses of hundreds and hundreds of primates. It's true for marsupials, for birds, for cetaceans, and it's even true for insects. If you look at different subcategories of butterflies that depend more or less on learning, what you

see is that they have a different developmental trajectory, such that the ones that depend on learning have a longer period of immaturity and produce fewer offspring. It turns out to even be true for plants and for immune systems.

Creatures that have more complex immune systems also have this longer developmental trajectory. It looks as if there's a general relationship between the very fact of childhood and the fact of intelligence. That might be informative if one of the things that we're trying to do is create artificial intelligences or understand artificial intelligences. In neuroscience, you see this pattern of development where you start out with this very plastic system with lots of local connection, and then you have a tipping point where that turns into a system that has fewer connections but much stronger, more long-distance connections. It isn't just a continuous process of development. So, you start out with a system that's very plastic but not very efficient, and that turns into a system that's very efficient and not very plastic and flexible.

It's interesting that that isn't an architecture that's typically been used in AI. But it's an architecture that biology seems to use over and over again to implement intelligent systems. One of the questions you could ask is, how come? Why would you see this relationship? Why would you see this characteristic neural architecture, especially for highly intelligent species? We're way out on the end of the distribution. Chimpanzee young are producing as much food as they're consuming by the time they're seven, but we humans aren't doing that even in forager cultures until we're fifteen, and we have much larger brains and much greater capacities for intelligence.

A good way of thinking about this strategy may be that it's a way of resolving the explore-exploit tradeoffs that you see all the time in AI. One of the problems that you have characteristically in AI is that as you get a greater range of solutions that seem to be moving in the direction of a system that's more intelligent, a system that understands the world in more different ways, what you also have is a big expansion of the search problem. If there are many more different things that you can do, how can you search through that space more effectively?

One way to solve that problem that comes out of computer science is to start out with a very wide-ranging exploration of the space, including parts that might turn out to be unprofitable, and then gradually narrow in on solutions that are going to be more effective. My slogan is that you could think about childhood as evolution's way of doing simulated annealing. It's evolution's way of starting out with a very high temperature, broad search and then narrowing it. The problem with a high temperature search is that you could be spending a lot of time considering solutions that aren't very effective, and if you're considering solutions that aren't effective, you aren't going to be very good at effectively acting in the world, performing the four Fs and doing all the other things that we need to do as adults.

An interesting consequence of this picture of what intelligence is like is that many things that seem to be bugs in childhood turn out to be features. Literally and metaphorically, one of the things about children is that they're noisy. They produce a lot of random variability. When I'm trying to explain the annealing idea to a general audience, I'll say, "There are two ways of thinking about this system. Here's a big box full of solutions, and you could be wildly bouncing around this box going from point to point and bouncing off the walls, or you could just be staying in one place and carefully exploring the space. Which one of those sounds like your four-year-old?" That randomness, variability, and noise—things that we often think of as bugs—could be features from the perspective of this exploratory space. Things like executive function or frontal control, which we typically think of as being a feature of adult intelligence—our ability to do things like inhibit, do long-term planning, keep our impulses down, have attentional focus—are features from the exploit perspective, but they could be bugs from the perspective of just trying to get as much information as you possibly can about the world around you.

Being impulsive and acting on the world a lot are good ways of getting more data. They're not very good ways of planning effectively on the world around you. This gives you a different picture about the kinds of things you should be looking for in intelligence. It means that some of the things that have been difficult for AI to do—like creativity, being able to get to solutions that are genuinely new and not crazy—are things that human children are remarkably good at. In our empirical evidence, they're often better at it than human adults are.

You can have a lot of random search, or you can solve a problem that's very highly constrained, but the combination of being able to solve problems that are highly constrained and search for solutions that are further away has been the most challenging problem for AI to solve. That's a problem that children characteristically solve more effectively than adults.

There are some other consequences of thinking about this particular life history as a solution to intelligence. For example, one of the things that we know children do is get into everything, and one of the things that we know that adult scientists do is experiment. That's active learning, where you're determining what your data sample is going to be and you're literally and metaphorically expending energy on getting the right kind of data sample, one that will not only be useful but will be the exact kind of data that will cause you to change the current view that you have of the world. It's a very unusual thing to be able to do, to go out into the world and spend calories and energy in order to turn out to be wrong. That's something that children very characteristically do, and if Danny Kahneman were here he could tell you adults very characteristically don't do.

Another aspect of what children are doing that would be informative for thinking about intelligence in general, is that children are cultural learners. One of the effects of this life history for human beings, in particular, is that it gives us this capacity for cultural ratcheting. It gives us a way of balancing innovation and imitation. If all we did as a result of cultural learning was imitate exactly the things that the previous generation had done, there would be no point in having cultural learning. There's constant tension between how much you are going to be able to build on the things that the previous generation has done and how much you are going to be producing something that's new enough so it would be worth having the next generation imitate. Having this developmental trajectory where you start out with a broad exploration and then narrow in on exploiting particular solutions gives you a way of solving that problem in the context of cultural evolution.

There are other ways that you can do that even as an adult, like have an interdisciplinary conference, or give adults things to do that are new. Recently, I've been interested in looking at psychedelic chemicals, which seem to have the rather surprising effect of putting adult brains back into a state of plasticity that looks much more like childhood brains. So, the effect

of psychedelic drugs neurally is that it increases the local connections and breaks the long-distance network connections. It literally induces plasticity and induces more synaptogenesis.

The ones that have been studied the most are psilocybin, LSD, MDMA, and ketamine, all of which have the same phenomenological properties. They also all turn out to have this same neural effect of driving the system back to something that looks more like childhood plasticity, which may be an interesting way of testing some of those ideas. It would be a good explanation for what otherwise seems very puzzling, which is that a small chemical change, at least by report, can lead people to have large changes in the ways that they see the universe or in the ways that they behave.

One of my slogans is that you could think about psychedelics as doing for the individual what childhood does for the culture: It takes a system that's relatively rigid and injects a bunch of noise and variability into the system, shakes it out of its local optima and lets it settle into something new.

Thinking about learning, in terms of active learning—having computers that would go out and play, and explore, and get into things the way that young children are playing and exploring and getting into things—is a sense in which children might be a model for intelligence that's different from the models of intelligence that we currently have. Thinking about systems that are learning from previous generations could be a model for intelligence that's different from the models of intelligence that we currently have.

Thinking in this life history perspective, another thing that's distinctive about human intelligence is that having a life history with a long childhood and a lot of caregiving changes our conceptions of moral relations. Our model for naturalizing morality has very much been a model of contracts, thanks to people like Robert. It's been a model of having individual people who are more or less equal in their status and in their relation, who are trying to develop a contract that will lead to the best outcomes for both of them. If you think about both markets and democracy, those are essentially wonderful institutions and inventions that maximize that process of contract-making so that we don't have to have face-to-face contracts to maximize our preferences.

If this picture is right, caregiving relations are absolutely key to having this life history. Every parent, no matter how bizarre or weird or crazy their child is, is committed to taking care of that child. That's a very different kind of relation than a contractual relation. It's asymmetric. Maybe your kids are going to take care of you when you're old, but it's not clear that they will, and that doesn't seem to be the motivation behind the life history. There's something about protecting a next generation that can introduce variability into the system. They have this fundamental asymmetry and transparency about them, so when you're attached to a baby, for example, it doesn't matter very much. You don't know very much about what the properties of the baby are. You don't know whether that baby is going to turn out to be valuable or not. There's just this transparent attachment that you have. It's having that transparent attachment that lets you have the noise and variability and mess.

If you were only attached to your children because you thought they were going to come out really well, or you wanted the children to come out as well as they possibly could, the sensible thing to do would be to look out at the universe of children and find the ones that you felt were most likely to succeed and then have everybody put all their love and attention into those and let the other ones perish. That seems like a crazy system. Part of the reason why it's a crazy system is because if you think children are the source of unpredictable variability all the time, then the moral commitment that you need to be able to allow unpredictable variability to thrive is not anticipating what the outcome of caring for that child. There's a lot of human moral and political life that has that character of unconditional commitment to a person, or to a community, or to a nation, and there's a puzzle about why those unconditional commitments give us a moral dimension that's different from the tit-for-tat contractual moral commitments.

One of the things that is fascinating about the Macy Conferences is that they have some of the earliest studies of things like longitudinal language acquisition before language development was a discipline within the official disciplines of psychology. People in that group were doing that kind of work, which echoes things that people here have said: If you want a good account of intelligence, thinking about those developmental trajectories both in the literal sense of thinking about children and adults, but also thinking about developmental trajectories more generally—thinking about developmental

trajectories over history, thinking about the ways that you could adjust to an environment over time—those are going to be a crucial piece of the story that's missing from the kinds of accounts that we typically have now.

* * * *

NEIL GERSHENFELD: There are beautiful algorithms emerging in machine learning that nicely interpolate between simulated annealing and gradient descent. You were describing those as extremes, but what these algorithms do is start by sampling the space, from there you use it to make an estimate of the distribution they were drawn from, then you re-synthesize from the distribution you estimated and use that to re-estimate the distribution. The way those propagate is they start looking like simulated annealing but they end up looking like gradient descent. In a nice way, the model grows. It might be an interesting analogy.

ROBERT AXELROD: The gradient descent is a function of temperature. So, if you have a high temperature, you're not doing gradient descent.

GERSHENFELD: What I'm describing is not simulated annealing. Simulated annealing is a simple thermodynamic model. Simulated annealing does a bad job of using local gradient information, which is the basis of back propagation in machine learning. What I'm describing is something that crosses over in an interesting way. You start by sampling a distribution broadly, then as you re-sample it you start to tighten the estimate, and then as you tighten the estimate you end up doing something that looks like gradient descent. One of the banes of simulated annealing is determining the cooling schedule and how to do the innovation. This is a very different way to answer it that crosses over between them.

AXELROD: I like the simulated annealing metaphor a lot, but what I was thinking is that you were describing it as, in simulated annealing terms, lowering the temperature. You broaden expiration and you lower the temperature. But simulated annealing itself typically raises and lowers the temperature on some schedule so you can jump out of the local optima. I was thinking, do children then become more plastic? Is adolescence is like that?

GOPNIK: One of the things that we have discovered empirically by looking at some of this is that if you look at physical problems, like trying to figure out how a machine works, what you see is something that looks like high flexibility and high search early on, and then it drops around school age and stays the same. Maybe it's that debate about whether that's the effect of school or the effect of school age. It's probably the effect of school age. It stays the same and then drops in adolescence. If you take a social problem, what happens is that you get the most flexibility in adolescence. Preschoolers are very flexible, then you decline. Adults are not very flexible. Adolescents are showing this peak that fits the neural evidence about plasticity, specifically in social areas. Presumably something like graduate school is a way of doing the same thing, plunking people into a situation in which they're forced into increased plasticity.

TOM GRIFFITHS: So, if I understand your argument, graduate school is like taking LSD?

GOPNIK: Yes, at its best. Sometimes those two things are combined, but the general idea would be that being put in a space in which the usual exploit strategies that you've learned are not effective has some of the same effects. That's true. The vividness of phenomenology, the vividness of experience, the emotional lability, which is characteristic of preschoolers, of people on psychedelics, of going to the center at Stanford for a year—those things are not just a joke, they're connected to one another.

I've talked to a lot of people who are doing machine learning, and what they typically say is, "Yes, we use annealing schedules all the time, but that's one of those artisanal things." There aren't general proofs about the way that the annealing schedule should work, about what's more effective in this context, about a general principle—aside from the general optimization idea, or the general getting out of a local minimum idea. Those things don't seem to be understood in a coherent theoretical way.

GERSHENFELD: They're not. I'll give you some references to these algorithms I'm mentioning in response to that.

SETH LLOYD: May I share *Wired*'s best scientific graphic of 2015 with you? This is from a paper by friends of mine. We're doing these quantum

algorithms for topological analysis of data. This is called "Homological Analysis of Brain Function." They took functional MRI data of the brain. The one on the left basically shows clusters of thought processes. There are about seven highly clustered processes, and they're talking between them with little links like this. Then the other one on the right is the same group of people having taken psilocybin. Let me just summarize if you haven't seen the picture of it. When you take psilocybin it's like, "Wow, everything is connected man."

GOPNIK: It is literally true that if you look at the developmental neuroscience literature, you essentially see that graph but going in the other direction. What you see is lots and lots of local connection. And this is boilerplate. One of the few things we know about developmental neuroscience is that you start out with lots of local connectivity and then as time goes on you get segregation. That graphic is part of why I'm making this argument.

RODNEY BROOKS: Getting back to life history, you were painting a very broad picture—humans and great apes— but other animals also have the characteristic that it's not that the group of children all happen at the same time, progress in the same period, and then go away. In human families over a period of time, there's a lot of sibling rivalry and learning from siblings. That happens in most great apes. Does it happen in other animals, too, or is that unique?

GOPNIK: One of the things that seems to be different is that in most animals you have a clutch of young all at once, so they're all on the same developmental progression.

If you look at the parental investment side, humans have pair bonding and alloparenting, including siblings being involved in care. The fact that you've got distributed siblings means that older siblings are involved in a lot of caregiving. They have postmenopausal grandmothers and of course they have biological mothers—it's those adaptations that you see in individual species. But I don't think there's any other species that has all three of them: pair bonding, alloparenting, and postmenopausal grandmothers and grandfathers. Well, grandfathers are more complicated because they're not postmenopausal, but you have this extra twenty years essentially that

people are investing in care. Not only do you have much more caregiving, but you have much more distributed caregiving, if part of the picture is supposed to be this picture of introducing a burst of noise, as it were, in each cultural generation. Part of getting that noise is just having these noisy children, but part of it is also the fact that very different people with different knowledge are giving them different kinds of information and models about what the culture is like.

GEORGE DYSON: Young killer whales are educated by their grandmothers.

GOPNIK: Yes, killer whales are a wonderful example of this. When I was talking about life history, I said we are also the only species that has postmenopausal grandmothers, except killer whales. Killer whales, go figure. It turns out that killer whales also have more culture than even other smart cetaceans, not just the adaptation to intelligence but the adaptation to culturally transmitted intelligence seems to be connected to this second-generation transmission.

G. DYSON: One of the grandmothers off Vancouver Island just died. She was 105.

GOPNIK: There's some pretty good evidence that because the young aren't dispersing for killer whales as they are with other cetaceans, the existence of the grandmother is changing the survival rates for the children and even for the grandchildren. So, when the grandmothers die, that affects the entire community.

There's good anthropological evidence that among human foragers things like myths, and songs and stories—things you might think of as giving you some of the high-level dimensions of what a culture has discovered—that transmission comes from grandparents to grandchildren. It skips parents in forager cultures, so parents are busy telling you what you should do specifically to hunt in a particular place, but the big ideas about what we've discovered about the world in general are coming from grandparents to grandchildren and skipping over the parents.

JOHN BROCKMAN: What if the grandparents are dead?

GOPNIK: Well, the older generation, in general. In forager cultures, it's going to be the fifty to seventy-year-olds. The comforting just-so story is that if you believe that, then remembering the things that happened yesterday if you're a grandparent is not going to be very useful because the kids already know that, and the parents can know that. Being able to talk a lot about the things that happened to you when you were very young, that's the stuff that you want if you're going to transmit information appropriately to children. As someone with an aging memory, I find this to be extremely comforting.

GERSHENFELD: In numerical methods, those algorithms lead to two diverging interspersed sets, we could argue. If you have these grandparents and those grandchildren, and then these grandparents and those grandchildren, but you get two interspersed sets. They begin to diverge and that, you could argue, might be why we have generations that alternate.

W. DANIEL HILLIS: Presumably, the annealing schedule for the human mind is optimized not just for the learning phase. We also have the role of being teachers and caretakers. For instance, it may be that it's better to turn off learning language when you're trying to teach a child language. That's Marvin Minsky's theory of why it got hard to learn languages when you're an adult.

The interesting thing with machine intelligences is that the modes of transmitting information might be completely different. In some sense, we've got a kludgy method of transferring knowledge from our mind into our children's minds. Certainly, with many representations, in machine knowledge there are much more efficient ways of doing that. In some sense, a machine can be born with all the experience of the previous generations of machines. I'm curious if you think that would radically change the annealing schedule of a machine.

GOPNIK: The proposal would be that a machine that was doing that without loss and without noise would be bad. What you want is for each generation, as you're getting the information from the previous machine, you'd also want to introduce a bunch of extra noise and variability.

HILLIS: That might be true, but I don't see why that follows. You don't have that option in the human method of transmitting knowledge because there's no mechanism by which you could transmit the knowledge through birth.

GOPNIK: We know something about some of the mechanisms of transmission, and there's this interesting debate in the cultural evolution community about this phenomenon called over-imitation. It seems to be very characteristically human that when we're imitating what another human does, we imitate even fine level details we don't need to imitate, things that aren't obviously relevant to the activity that the person is performing. You can take chimps and children and have someone perform a whole bunch of complicated bells and whistles to bring about a particular kind of effect, and the chimps will read through to what the actual problem is that you're trying to solve, but the kids will put in the bells and whistles. Presumably, computers could do both, so the next generation of computers could simply take all the details about what the previous generation had done, but you'd end up with overfitting problems. That's a classic overfitting problem.

HILLIS: Wouldn't it be equivalent to just having somebody with a lot more experience and a lot more cases that they would learn from? That doesn't necessarily mean you're overfit. That's a different issue.

GOPNIK: Again, this is where it would be nice to have people working out the computer science to explain what you would expect to have happen in those conditions versus other conditions.

HILLIS: But my gut feeling is with more information, it would be better.

GOPNIK: Well, I'm not sure that that's true. Again, what might be happening is that having more information is just going to narrow the space of new solutions that you're going to search. Right?

BROOKS: The world has changed.

GOPNIK: Yeah, exactly. The world is changing.

HILLIS: Well, okay, that is another issue, but you can know that the world has changed, too, so you can weight them with time or something like that.

GOPNIK: I just saw a really wonderful paper about this is. There's good evidence, in birds for example, that environmental variability is a trigger for these life history changes, especially environmental variability within the lifetime of the organism, which seems to be the thing that triggers a long life history versus a shorter life history.

HILLIS: Put another way, I don't doubt that different information has different amounts of relevance. It seems unlikely to me that the information that was available to you right from the moment you were born happens to be just exactly the best set of information. It's much more likely that it's being able to choose from a wider set of information and weighting it appropriately, which is not an option with human children.

GOPNIK: Why is it not an option in human children?

HILLIS: Because the experiences weren't recorded in a way that they can in some sense retrain on them.

CAROLINE JONES: But they're out having experiences. They're out in the world having experiences. That was your big box metaphor. They're not constrained to an information transfer from the parents. There are also agents in the world.

PETER GALISON: It's interesting to look at what happens within a discipline like physics where you can have a group of people, like the people who formed quantum mechanics—Niels Bohr and Heisenberg—and one of the things that they did over the course of their life was to come back over and over again to the hope that an extreme excursion from what was known was what they needed. For example, wanting to give up the conservation of energy, as Bohr did on two or three different occasions. In 1935, Heisenberg said we need a new revolution to understand why some things that look like electrons could penetrate a lot of lead and others couldn't. It turned out you just had to stick with the physics they knew and work it out, and it turned out that there's just a heavier version of the electron.

Heisenberg thought there were revolutions all the way up past the war, and he got the young German physicists after World War II into a whole mess of trouble because they were departing from productive physics. I just say this

because if the opposite of a trauma is a traum, they had this dream experience as young people, Bohr in 1913 and then later, and Heisenberg when he was practically a kid in 1924, ’25, ’26. Then they kept looking for that again over and over again.

You can have the consequences of overconservatism being growing old and not being willing to meet new ideas in some way to have a high enough temperature of excursion in the annealing process, but if you design the computer that was always making huge excursions, you’d be in a world of hurt intellectually. One of the problems is, how do you know contextually whether it’s time for a high temperature or a low temperature?

GOPNIK: This is relevant to what Robert was talking about: Exactly how do you balance those things across a scientific discipline? In a way, evolution gives it to us for free with childhood because children aren't sitting there saying, "In this context, should we be exploratory or not? Is this insane imagined fantasy going to turn out to be useful in the long run or not?" They just do it. That’s just the way that they’re designed. When you have social institutions that are trying to do the same thing, trying to balance those things, or when you’re trying to design a computer algorithm, then the question about whether there are contextual cues that you could use gets to be a relevant problem. There’s a little bit of work in the developmental area about these "live fast, die young" life history strategies, even within a species, versus having a long extended exploratory period strategy. There’s a lot of debate. It’s not obvious.

One thing is when the environment is variable in particular kinds of ways over particular time scales, it’s an advantage to explore. It looks as if when you’ve got a lot of resources, then you can afford to have more exploration, which you can’t when you have fewer resources. There’s some evidence that kids who are under stress or maturing animals that are under stress mature more quickly. That’s also underresearched, and the intuitions that you have don’t necessarily translate into what happens when you do the math.

DAVID CHALMERS: This is super domain relative. There are critical periods for language learning, early, and then for music appreciation much later, like when you’re eighteen or something. So, the annealing has to be domain

relative. I guess what I’m wondering is whether there are domains where kids are super-conservative, non-exploratory.

GOPNIK: Kids have a single utility theory, which is, "Be as cute as you possibly can be," and they’re extremely good at maximizing that utility. No other utility function is relevant to you if you’re a kid, but it turns out that being as cute as you possibly can be is not trivial. Having a caregiver environment that’s highly stable and predictable, when you don’t have to do any cognitive work in terms of wondering whether you’re going to be taken care of or not, that’s something that’s not transparent or easy. That’s a context where children are extremely conservative. When it comes to their parents, they don’t want variability. They don’t want change. They don’t want to noise. They’re very conservative about that.

BROCKMAN: We’ll leave it with "Be as cute as you can possibly be."

TOM GRIFFITHS
Humans: Doing More with Less

Imagine a superintelligent system with far more computational resources than us mere humans that's trying to make inferences about what the humans who are surrounding it—which it thinks of as cute little pets—are trying to achieve so that it is then able to act in a way that is consistent with what those human beings might want. That system needs to be able to simulate what an agent with greater constraints on its cognitive resources should be doing, and it should be able to make inferences, like the fact that we're not able to calculate the zeros of the Riemann zeta function or discover a cure for cancer. It doesn't mean we're not interested in those things; it's just a consequence of the cognitive limitations that we have.

As a parent of two small children, a problem that I face all the time is trying to figure out what my kids want—kids who are operating in an entirely different mode of computation, and having to build a kind of internal model of how a toddler's mind works such that it's possible to unravel that and work out that there's a particular motivation for the very strange pattern of actions that they're taking.

Both from the perspective of understanding human cognition and from the perspective of being able to build AI systems that can understand human cognition, it's desirable for us to have a better model of how rational agents should act if those rational agents have limited cognitive resources. That's something I've been working on for the last few years. We have an approach to thinking about this that we call resource rationality. And this is closely related to similar ideas that are being proposed in the artificial intelligence literature. One of these ideas is the notion of bounded optimality, proposed by Stuart Russell.

TOM GRIFFITHS is the Henry R. Luce Professor of Information, Technology, Consciousness, and Culture at Princeton University. He is co-author (with Brian Christian) of *Algorithms to Live By.*

* * * *

TOM GRIFFITHS: I'm going to talk about two problems that seem contradictory, but I'm going to argue how they are intimately related to one another. The first problem is that people are still smarter than machines. This is not necessarily a problem for people; it's more of a problem for machines. Despite the recent advances in AI, you can point to lots of individual things that people can still do better than computers can, but, more generally, you only have one system that is capable of doing all of those different kinds of things, and that system is human beings.

The current trend in machine learning is one of solving problems by increasing the amount of data and the amount of computation that get thrown at them. If I were showing slides here, I would show you a nice picture that some of the people at OpenAI made, where they took a bunch of the recent milestones in AI, starting from image net classification through things like AlphaGo and AlphaZero, and they plotted out as a function of time how much compute went into each of those things. You'd see there's a nice increasing line. I would argue that focusing on that trajectory is something that isn't necessarily going to take us in the direction of getting systems that can do the kinds of things people can do, particularly, this generality that characterizes human intelligence.

As a historical example, the interaction between Deep Blue and Gary Kasparov has been taken as evidence for the success of AI, but you can instead look at it as revealing something important about the power of human cognition. While Deep Blue won the majority of those games, they were doing it under entirely different conditions. Kasparov was playing with the energy equivalent of a light bulb and was able to evaluate maybe three different moves a second, whereas Deep Blue was playing with a huge amount of energy resources going into it and the capacity to evaluate something like 100,000 moves per second.

The critical difference there is that one of the things that helps to make human beings intelligent in the way that human beings are intelligent is intrinsically the fact that we have limited cognitive resources. Our ability to efficiently manage, use, and deploy those cognitive resources in different ways to engage with the different kinds of computational problems that we encounter is part of what makes us intelligent in the way that we characteristically think is intelligent.

If we're doing slogans like Alison Gopnik was doing, the slogan here would be, "Humans: Doing More With Less." That kind of perspective is not necessarily one that is encouraged in the current machine-learning based approach to AI, but it's going to be critical to being able to succeed in getting past some of the challenges that the field is currently facing.

The second problem, which as I said seems at odds with this, is that people are not so smart. On the one hand we have people smarter than machines, and on the other hand we have people with a reputation for being dumb. You have heard hints about this reputation. People have well known cognitive defects, and Danny Kahneman is one of the people who helped to reveal those defects. The way in which those defects are typically characterized is in terms of a comparison of human beings against a classical notion of rationality. This classical notion of rationality isn't a good criterion for evaluating human behavior or, importantly, the behavior of machines.

This classical notion of rationality says that in any situation what you should be doing is taking the action that maximizes your expected utility without regard for how hard it is to compute that action. That characterization of rational behavior is something that is not achievable by any realistic organism, whether it be a human being or a computer, because all realistic organisms are limited in the amount of computation that they have available to them.

The reason why this is something that matters in the context of both AI and understanding human cognition is that it suggests there might be a different way that we could go about characterizing what constitutes rational behavior for realistic entities. It might be one that gives us different insight into understanding the ways in which human beings behave, and whether or not the things that we do constitute cognitive defects.

Part of the reason why that's important for AI is not just because that's a criterion that we're going to hold AI systems to, but because if it gives us a model of human behavior that has the same generality as that classic notion of rationality, then it gives us an important tool that AI systems are going to need in order to be able to act in ways that are beneficial to humans. It gives us a component of a system that is going to be able to make

inferences about what human beings want based on the ways in which human beings behave.

Imagine a superintelligent system with far more computational resources than us mere humans that's trying to make inferences about what the humans who are surrounding it—which it thinks of as cute little pets—are trying to achieve so that it is then able to act in a way that is consistent with what those human beings might want. That system needs to be able to simulate what an agent with greater constraints on its cognitive resources should be doing, and it should be able to make inferences, like the fact that we're not able to calculate the zeros of the Riemann zeta function or discover a cure for cancer. It doesn't mean we're not interested in those things; it's just a consequence of the cognitive limitations that we have.

As a parent of two small children, a problem that I face all the time is trying to figure out what my kids want—kids who are operating in an entirely different mode of computation, and having to build a kind of internal model of how a toddler's mind works such that it's possible to unravel that and work out that there's a particular motivation for the very strange pattern of actions that they're taking.

Both from the perspective of understanding human cognition and from the perspective of being able to build AI systems that can understand human cognition, it's desirable for us to have a better model of how rational agents should act if those rational agents have limited cognitive resources. That's something I've been working on for the last few years. We have an approach to thinking about this that we call resource rationality. And this is closely related to similar ideas that are being proposed in the artificial intelligence literature. One of these ideas is the notion of bounded optimality, proposed by Stuart Russell.

Basically, what we want to do is come up with a criterion that describes how a rational agent, be it a human or a computer, with limited computational resources should use those computational resources and then act. You can think about this as characterizing a kind of optimization problem, similar to the classical optimization problem which says that what you want to do is maximize your expected utility. We're going to think about how to go about choosing an algorithm that is going to lead to an action we take which

maximizes expected utility while minimizing the associated computational costs.

If you have a model of the computations that are available to an agent, the costs that are associated with those computations—the amount of time that they take or amount of other kinds of resources—then you can define an optimization problem, which then gives us a way of saying what constitutes rational behavior. Rational behavior is no longer the agent who always takes the perfect action in the perfect circumstance, it's the agent who follows the algorithm that leads them to take the action that best optimizes this joint criterion of maximizing expected utility while minimizing computational cost.

What I want to do is give you one concrete example of a way in which that is useful in understanding one of these classic cases where people behave irrationally. And that example is what's known as the availability heuristic, particularly, the overrepresentation of extreme events. One way in which people often act irrationally with respect to a classical criterion is that if you ask them to estimate the probability of something like a terrorist attack, or a shark attack, or these other extreme negative circumstances, they significantly overestimate those probabilities.

So, when you're getting on a plane, you're spending more time than you should thinking about the possibility that the plane will crash. When you're making a decision about going snorkeling, you're thinking not enough about the tiny worms on the corner of the coral reef, and far too much about the sharks that are very unlikely to bite you. Those things seem irrational. They're things that are going to affect your behavior in ways that aren't necessarily consistent with the way in which a purely unbiased agent who's appropriately evaluating expected utilities might act.

When we put that in the context of a resource rationality framework, we need to define the computational problem that we want to solve, talk about the resources that we have available, and then think about what the best kind of strategy is for deploying those resources. In this context, it's clear that what we want to do is evaluate something like an expected utility. That requires summing over all of the possible outcomes of our action—the utility of that outcome, multiplied by its probability. That's potentially a costly

procedure. When you're taking an action in the real world, there are many possible outcomes and they're going to have a variety of different utilities.

Let's say you were going to try and approximate that calculation. And the way you're going to try and approximate it is by drawing samples, doing a Monte Carlo approximation. You're going to sample some possible outcomes, and then you're going to consider the utilities of those possible outcomes, add those up, and that's going to be the way in which you're going to evaluate whether you should take that action. The choice that you have to make is a choice about what distribution you're going to sample from. Your goal is to choose a distribution such that you're able to draw a relatively small number of samples because those samples are costly. That's time you're spending standing around rather than going snorkeling. You want to come up with a way of drawing those samples that will allow you to minimize those costs, make decisions quickly, make decisions with small numbers of samples.

When we think about trying to do an approximation via this Monte Carlo procedure, the intuitive, straightforward thing people think of is to sample directly from the distribution that you care about. You're going to think about possible outcomes, but you're going to think about them with the probabilities that are associated with those outcomes. So, you sample from that distribution that has the benefit that it gives you an unbiased estimate of the probabilities. But if you're in a situation where there's an extreme distribution, a skewed distribution of possible utilities, and where there are low probability events that have extreme negative utility, that strategy doesn't work very well. The reason why is that the estimate that you end up getting is one that has a huge amount of variance. So, from one set of samples to another, there can be a very big difference in the value that you get because you may or may not have included those extreme events.

If we're dealing with small samples, it's not the bias in the estimate that's going to be the problem, it's literally the variance in the estimate that is going to end up killing us. So if you were going to make a decision about whether you should play a game of Russian roulette and there is a revolver in front of you with six places for bullets, one of which is actually a bullet, you can calculate how many samples you need to draw in order to be 99.9 percent sure that you should not play this game. It's something like 51

samples. There’s a lot of variability there. You want to be able to make that decision much more quickly.

So, what’s the distribution from which we should sample in order to minimize that variance? There's a nice result which says that that distribution is proportional to the probability of an event occurring multiplied by the absolute value of the utility of that outcome. And this results in a biased estimate. It’s biased in the direction that it’s going to overestimate the probability of extreme events, but it’s the estimate that reduces variance. If you’re trying to do the least computation you can do while minimizing the chance that you end up accidentally killing yourself, this turns out to be the best strategy. It’s a resource rational strategy.

So, what you do is you wander around the world, you encounter those events and you remember those events as you encounter them. But the probability that you remember them or the probability that you retrieve them from memory is proportional to the absolute value of their utility. That’s a mechanism that instantiates that kind of stuff.

Herb Simon talked about this idea of bounded rationality, but he was very reluctant to define what bounded rationality meant. And in fact, there’s a letter that he wrote to Gerd Gigerenzer in which he makes it very clear that that ambiguity was a feature not a bug, that it was intended to make people think about alternatives. The way that I think about it is that the notion of boundedness picked out a subset of the space of possible strategies that you can follow. That’s the optimal strategy. That’s the thing that’s picked out by the classical notion rationality. Bounded rationality says you don’t have the resources to get there, so now there’s a space of alternatives. And then bounded optimality says, out of that space you’re going to be choosing the thing that is going to be the best thing. It gives you back that optimization criterion that gives you a way of then having a theory of behavior which has the generality of that classical notion of rational behavior.

The point of this example is that this is a case where it seems irrational under a classical notion of rationality, but it makes sense under a more realistic characterization of what rational behavior might be like. There are other cases that we work through where we can show that some of these classic heuristics fall out of this approach. And while heuristics result in

biases, being biased doesn’t mean that you’re not doing the thing that is resource rational. Bias is a natural tradeoff to accept in order to allow you to operate with limited cognitive resources. Just because you see biased behavior, doesn’t mean that people aren’t doing something that makes sense. It might mean that people are doing something that makes sense, but they’re operating under resource constraints.

This ties back to some of these questions about machine learning that I started out, the immediate question that should come up for you is if people are following these resource rational strategies, how are they discovering them? How are we ending up finding good ways of using the limited cognitive resources that we have? This is a problem that’s called rational meta-reasoning. How do we rationally reason about the strategies that we should be following as agents in terms of the computational resources that we deploy for solving particular problems? We're not just reasoning, we’re reasoning about the way in which we should reason. Again, it's something where you can do a little bit of work and get a lot of leverage.

There are a couple problems with rational meta-reasoning. One is a problem that we call algorithm selection. This is a case where you know what the algorithms are and you’re trying to choose between them. The more interesting case is what we call algorithm discovery, where your goal is to come up with the right algorithm to use to solve a particular problem, putting together these pieces of computation that you’re going to use to solve that problem.

The way that you can approach this is by recognizing that in fact this deployment of an algorithm is itself a decision problem. It’s a sequential decision problem where you’re making a decision about the sequence of computations that you’re going to execute, one after another. Framing it in those terms allows us to tie it back to classical problems that are faced in decision-making, reinforcement learning. You can characterize the problem of selecting what computation to perform through this process as a Markov decision process, which is something that we can solve using these classical tools.

One of the ways in which most machine-learning methods work at the moment is, you take a single monolithic neural network and you throw it at

your problem. That's different from the way in which we've been characterizing human cognition as working, which is that you've got the pieces of these computations and you're deciding how to put these together to solve different problems. And then in the neural network world, that's about constructing what they call a computation graph, the sequence of transformations that you're applying in order to get to a solution. So, you can formulate the problem of constructing the right neural network to use to solve a particular problem as a decision problem in these same terms. And it's a decision problem about the computations that you'd apply in order to solve a problem in the world.

Using these kinds of tools and engaging with a set of problems in ways that are much broader than the canonical approaches currently being used in machine learning, such as trying to train the same system to find the right cognitive modules to put together in order to solve all of these different problems, gives us a path beyond the current monolithic approach to building these machine-learning systems, and maybe a path towards building more human like AI systems.

* * * *

JOHN BROCKMAN: I have a question vis-à-vis the claims made by the deep-learning community. Where is this going? When you talk about the next revolution, is this going to be AI that we can use?

GRIFFITHS: I'm focused on what I see as the gap between what people do and what current approaches in machine learning do. Whether you want to close that gap is going to depend on what your motives are. If you just want to build the best image classification system, then you don't have to care about this. But if you want to build a system that is doing a wide range of tasks in the same general way that humans do, then those are the kinds of architectural questions that you want to ask.

ROBERT AXELROD: How would you deal with the idea that humans eat too much sugar and fat? We know the evolutionary reasons why, but the human is told that it's not optimal for your health and they do it anyway. That is not a question of limited calculation, but it is sub-optimal behavior.

GRIFFITHS: I'm not claiming that all instances of sub-optimal behavior are explained in this way. Adopting this perspective changes the way that you should think about debiasing. Classical approaches to debiasing try to make people act in ways that are more rational. But from my perspective that's not going to work, because the strategy that people are following might be a good strategy given the resources that are available to them, so instead of focusing on modifying people's strategies, what you do is focus on modifying the structure of the environment around them.

AXELROD: You could also modify their resources. For example, if you're teaching calculus, they can think about rates of change in a way that somebody who is naïve cannot.

GRIFFITHS: You can change the set of computations that they're able to perform, and as a consequence that can change the strategies.

PETER GALISON: When you were talking about bounded rationality, we often assume that if we don't have the resources to sample very widely, we're making poorer decisions than we would have if we could have sampled more widely, but we understand it, because in that circumstance we didn't have access. So, being scared of a shark attack, we don't have that information, so we would have realized only a handful of shark attacks in ten years.

Recently, some of my nearest and dearest went swimming at the beach near where we live, where there was a shark attack two days before, and there were seals swimming there. The conditional probability was elevated, partly because there was a bound on accessed information. So sometimes it seems to me that bounded rationality could be better than the universal knowledge because it might include conditional probabilities that were specific to the circumstance.

So, it may be that people are not only being rational when they make these local decisions because their knowledge of the broader universe of data is limited, it may also be because they see something. They add factors that are relevant, or that they suspect might be relevant to their decisions that would make it quite sensible.

GRIFFITHS: Yes. I can give you an interesting example like this, which is we looked at a case in which you're doing decision making over some horizon, but we assume your horizon is limited relative to the problem that you want to solve.

So, say you're making decisions about a twenty-year trajectory in terms of your career, but you're only able to see or plan out into the future five years or something like that. Under those circumstances, it turns out that it's beneficial to be optimistic. So, it's beneficial to conflate your utility and probability in exactly the way that I'm talking about, where you over-estimate the probability in this case of good events.

And it's beneficial to be optimistic because if you were only able to plan over that limited horizon and you were being perfectly calibrated to what the probability of events are, you miss the chance to pursue a low probability outcome which has a big payoff in terms of fifteen years beyond the part which you're able to see. Whereas if you're optimistic, you erroneously pursue those attractive career opportunities that you're going to fail at, but the fact that you pursued them put you in the position to be able to then benefit from those in the future.

SETH LLOYD: I'm a little confused about what you mean by sampling here. I never quite understood what people meant by sampling. If you mean look at my past history and count the number of times I was attacked by a shark, then I estimate the probability to be zero. And that's clearly wrong, because you know that people are attacked by sharks. So, your information is just, "I read in the newspaper that somebody was attacked by sharks."

GRIFFITHS: When we talk about sampling as a cognitive mechanism, we're pretty agnostic about what the distributions are that you have access to.

LLOYD: Well, what do you actually mean by sampling? Do you mean I am drawing from some set of events with this distribution? How do you draw from that if you're talking about shark attacks?

GRIFFITHS: The best example of this is cases where you're able to somehow generate samples from memory, as a consequence of your experience, plus as you're saying the testimony of others and the things that

you've read about. And you're putting all of those together into a distribution that you're then generating samples from when you're considering outcomes.

FREEMAN DYSON: Is this just simply a formalized version of Kahneman's *Thinking, Fast and Slow*?

GRIFFITHS: What psychology has done very well, and what Kahneman's work is a good example of, is characterizing a wide range of circumstances where people behave in ways that deviate from this classical notion of rationality. What psychology has not done particularly well is develop a formal theory with the same generality as that classical theory of rationality that actually explains in particular circumstances how people are going to act.

F. DYSON: So, your answer is yes?

GRIFFITHS: My answer is, it would be very nice if we were able to formally express those things. And that's the kind of goal that we have in mind, yes.

ALISON GOPNIK: So, it's doing prospect theory right, right? The point is that prospect theory was supposed to be here's the positive theory that's the counterbalance to that.

GRIFFITHS: Prospect theory is still a descriptive theory that says that people are making decisions using this function to characterize probability, this function to characterize utilities, those are empirically derived functions. So, the kind of thing that you want to be able to do is to say we can derive from this why people are taking actions in this particular way. And I'll say that the method that I talked about, the sampling-based method, that actually predicts people's decisions in an empirical choice prediction setting better than prospect theory does. So that's just starting from first principles and driving the solutions to those problems.

GOPNIK: It is true that this gives you a good explanation of why people are acting the way they do, but another respect in which people seem to be superior to AIs gets back to this point about getting at vertical representations of what's going on in the world around us, notably in

science. Is that bug of having limited computational bounds actually a feature when it comes to extracting the structure of the world around us?

GRIFFITHS: It's a feature because it forces us to be good at this kind of metacognition. If you think about the kinds of things that AI systems currently struggle with, one of these examples is being able to find a reasonable sub-goal in a reinforcement-learning task. If you look at the set of computer games that AI systems can play better than people, and then you look at the ones where they fail, the ones where they fail are the ones where you have to formulate some kind of abstract goal, like "I'm going to get the key so I can open the door, which is going to appear on two screens."

But the human ability to do that is entirely a consequence of the fact that we have limited computation. If you were able to see an arbitrary distance into the future, you don't need to formulate sub-goals. You just follow the optimal policy. It doesn't need any decomposition to do that. Decomposing the problem in that way is what you do when you try and solve it with less computation. So, being able to form those abstractions allows you to represent the problem in a way where you're able to solve it, even though you're only able to consider three moves a second.

IAN MCEWAN: Are you saying what seems like cognitive defects are actually useful features because we're all descended from people who weren't eaten by sharks? We must be getting something right.

GRIFFITHS: I'd say what seemed like cognitive defects are defects in the sense that they're a consequence of the limitations that we operate under. But the fact that we operate under limitations means that we've had to develop the kind of cognition which is not well represented in current AI systems, which is being able to reason about how to use the bit of smarts that we've got to solve a wider range of problems.

MCEWAN: There seems to be something interestingly parallel with Alison Gopnik's talk of a push to humanize AIs. To get them to do something more like what we do, if they're going to live among us.

GRIFFITHS: That's another problem. It's important that we have a theory of how they work in order for us to interact with them, but it's even more important that they have a theory of how we work in order for them to interact with us.

One of the surprising things that we discovered is that expected utility theory is a terrible model of how people act, but it's a good model of how people think other people are going to act. We have a theory of mind, but our theory of mind is flawed in that we think that people are more rational than they are. If you wanted to make a machine that could reason well about how to interpret human actions, it would be nice if we were able to do so in a way that took into account cognitive limitations as well.

MCEWAN: In the model of my Adam, he understands humans by reading world literature. And since he's got a very good memory, the totality of world literature, imaginative literature, is not a prescription but a description of humans are. It doesn't necessarily provide solutions, it just takes us through all the moral corners that people have been able to imagine.

GRIFFITHS: Yes. And that's pretty consistent with the way that current machine-learning methods work, which is that you drown them in a huge amount of data and they're able to memorize the relevant aspects of the data and generalize, but they're not necessarily forming a systematic theory that they're able to use to generalize to new circumstances.

GOPNIK: There are reasons why his reading world literature as opposed to just taking in all the things that are in the papers might be an advantage. If you think about trying to see what the boundary cases are, so you're trying to figure out the structure of a particular theory, trying to figure out what the consequences of it are, you're often better off thinking about non-existent boundary cases than you are thinking about the things that you see all the time.

If you think about Einstein trying to explain his theory, if he just said, "Well, look, here's a prediction of the theory. If you drop this, it will fall at this particular speed according to my theory." That would not be very informative. Having these fictional boundary cases seems to be a better way

of telling you, "Here's the big important differences between my theory and other theories."

If you're thinking about human beings, for instance, it might be that having these fictional extreme boundary cases, which is what you typically find in world literature, is a better way of knowing what people's psychological structure is, or at least what people's theory of their own psychological structure is than it would be if you just looked at all the things that were in the newspapers.

MCEWAN: The sum of all the things that didn't happen is near infinite. The possibility for world literature as a mental space is infinite.

LLOYD: Speaking from my experience teaching the "Miracle Methods of Probabilities" to undergraduate mechanical engineers, which is what they learn, everything obeys a central limit theorem, the deviations are ne Gaussian. In a Six Sigma event, like in Six Sigma management, Six Sigma means it has a probability of one in ten to the 12^{th} of occurring. But from experience, we know that distributions have fat tails. There are power laws. They are lognormal distributions. And Six Sigma events occur all the time. If you designed a bridge thinking everything is just a Gaussian distribution and Six Sigma's never going to occur, then you're going to have bridges that are falling down right and left. I'm actually just agreeing with you. The fact that shark attacks occur, even though they're extremely rare, means you probably want to give it some thought that it might happen.

W. DANIEL HILLIS: It seems to me like this notion of bounded computation is also relevant to the discussion, early in the morning of Alison Gopnik's point about the complexity of goals and the insolvability of goals. I'm very bad at predicting what's going to make me happy or not sad, so I have goals instead. And what goals are, is they're basically admissions of failure of my ability to make bad computations. So instead what I try to do is act in a way toward doing something else that's a surrogate for that, a stand-in for that. And in the same way with machines, when I give machines goals, I can't really decide what's going to make me happy for the machine to do, so what I'm doing is I'm using my bounded rationality and establishing a stand-in for that of what would make me happy, for the machine.

GRIFFITHS: We've been thinking about ways of helping bounded humans do a better job of achieving their goals if they're able to specify what they are. It's based on reversing that loop. Instead of having humans define the reward functions for machines, you have machines define the reward functions for humans. We have an approach called optimal gamification. The idea is that you have a sequential problem you want to solve, and we can write it down, but you're not able to see far enough into the future to work out what the optional policy is, so you can give it to a computer, get the computer to solve that problem, and then we can take the solution that's computed by the computer and use that to construct and modify the reward function for humans, such that even perfectly myopic humans following the modified reward function will achieve their long term goals.

We have deployed this system which has shown that we can reduce people's procrastination, procrastination being a classic example where you've got a big payoff that's far into the future, and then the optimal modification of that reward function makes that long term payoff smaller but spreads it out through time so you can follow the bread crumbs and then eventually get there.

HILLIS: This also suggests a modification of the voting algorithm that you suggested earlier, which is the real great democracy would not be the one where you submitted an algorithm showing what your preferences were, but merely you submitted an algorithm that showed what would make you happy and then you, the voting commission, takes all those algorithms, runs them under different scenarios and picks the optimal one.

NEIL GERSHENFELD: I would caution about the three moves per second being misleading. If you view that as an update rate, it's updating a very high dimensional feature vector. Unlike the move generator you're comparing in Deep Blue. And so, if you look at high dimensional optimization algorithms, the effect of operations per cycle is a huge number because you're moving this giant feature vector and the three moves per second is the velocity of this high dimensional feature vector, which is way more than three moves a second.

FRANK WILCZEK
Ecology of Intelligence

I don't think a singularity is imminent, although there has been quite a bit of talk about it. I don't think the prospect of artificial intelligence outstripping human intelligence is imminent because the engineering substrate just isn't there, and I don't see the immediate prospects of getting there. I haven't said much about quantum computing, other people will, but if you're waiting for quantum computing to create a singularity, you're misguided. That crossover, fortunately, will take decades, if not centuries.

There's this tremendous drive for intelligence, but there will be a long period of coexistence in which there will be an ecology of intelligence. Humans will become enhanced in different ways and relatively trivial ways with smartphones and access to the Internet, but also the integration will become more intimate as time goes on. Younger people who interact with these devices from childhood will be cyborgs from the very beginning. They will think in different ways than current adults do.

FRANK WILCZEK is the Herman Feshbach Professor of Physics at MIT, recipient of the 2004 Nobel Prize in physics, and author of *A Beautiful Question: Finding Nature's Deep Design*.

* * * *

FRANK WILCZEK: I'm a theoretical physicist, but I'm going to be talking about the future of mind and intelligence. It's not entirely inappropriate to do that because physical platforms are absolutely a fundamental consideration in the future of mind and intelligence. I would think it's fair to say that the continued success of Moore's law has been absolutely central to all of the developments in artificial intelligence and the evolution of machines and machine learning, at least as much as any cleverness in algorithms.

First I'll talk about the in-principle advantages of artificial intelligence with existing engineering principles. Then I will talk about the enormous lead that

natural intelligence in the world has, although there are obviously great motivations for having general-purpose artificial intelligence—servants, or soldiers, or other useful kinds of objects that are not out there. Then I'll talk a little bit about the forces that will drive towards intelligence. Perhaps that's superfluous here, but we've been talking about how improvements in intelligence are an end in themselves, but it's worth at least saying why that's going to happen. Finally, I'll argue for an emphasis on a new form of engineering that is not being vigorously cultivated, and I'll draw some consequences for what the future of intelligence will be.

One of the advantages of artificial over natural intelligence is that they're extraordinarily powerful quantitatively and qualitatively. Take speed, for instance. Transistors, which are the basic decision-making processes or information processors in modern computers, operate at 10 billion operations per second. If you were to ask how fast human brains notice that movies are a series of still images rather than a continuous image, it's about 40 per second. There's a factor of a billion there, at least, plus an order of magnitude. Machines are a lot faster. They have much better error freedom and ability to correct errors. They operate digitally. Associated with that, they have the ability to download enormous amounts of information seamlessly and automatically.

Their architecture is known because they were built, so they're modular. You can add abilities to them, you can add programs, but you can also add senses. If you want them to, say, look at scenes in ultraviolet, you'd plug in an ultraviolet camera. They're ready for quantum mechanics, so if quantum mechanics turns out to be an important way of processing information because it opens up new levels of parallel processing, then, again, you can plug it in as a module. And they have a very good duty cycle. They don't need care and feeding and, most importantly, they don't die.

Artificial intelligence has many advantages, so it's almost paradoxical as to why they aren't doing better than they are. What advantages does natural intelligence have in the present competition? For one thing, it's much more compact. It makes use of all three dimensions, whereas existing semiconductor technology is basically two-dimensional. It's self-repairing, whereas chips are very delicate and have to be made in expensive clean rooms. Lots of things can go wrong with artificial intelligence, and errors

frequently make it necessary to shut down and reboot. Brains aren't that way.

We have integrated input and output facilities—eyes, ears, and so forth—that have been sculpted over millions or billions of years of evolution to match the world we find ourselves in. We also have good muscular control of our bodies and speech. So, we have very good input and output facilities that are seamlessly integrated into our information processing. While impressive, those things are not at all outside the plausible domain of near future engineering. We know how to make things more three-dimensional. We know how to work around defects and maybe make some self-repair. There are clear ways forward in all those things, and there are also clear ways forward in making better input and output modules.

Although the input and output modules for human brains are very impressive, they by no means approach physical limits. Even your intelligent phone can make better images and computers can talk. In some very restricted areas there are physical limits, but we don't exhaust physical limits, except in a few very exceptional cases. For instance, our resolution in space and time of vision, which is our best sense, is not that good. It only samples a limited part of the spectrum and even in that limited part of the spectrum takes three crude averages. We don't sense polarization. Machines can do all those things.

Where humans do have a qualitative advantage—far beyond anything in existing engineering—is in the connectivity and development of their basic units. The brain is made out of tens, or tens of tens of billions of units, each of which is an impressive module. Then there are the glia that help along. These were made by processes of self-reproduction and exponential growth. Current engineering doesn't have anything like that, where you have exponential growth of sophisticated units that self-reproduce. Brains also have enormous amounts of connectivity. Semiconductor technology has maybe a hundred connections per unit, whereas the brain has thousands.

These differences are so vast quantitatively that they count as being qualitative differences between current artificial intelligence engineering and natural intelligence. This is where natural intelligence has a big edge. And it gives a big utility.

I was very pleased to hear Alison’s talk first, because this touches on the learning algorithms and the learning process that humans use. They have this vast collection of neurons and connections and spend a lot of time getting rid of them and sculpting them. That’s the way human learning mainly works—by interacting with the world and getting feedback. Some connections get reinforced, while others get winnowed away. This has been discovered now to be a very powerful way of learning things in artificial neural nets. Real neural nets, however, are on another scale altogether because they’re bigger, better hooked up to the external world, and more connected.

Now I'd like to talk about why I think there will be an evolutionary drive towards increasing intelligence as an end in itself, a demand side as opposed to people who just want to make it better from a supply side. First of all, there are consumers. Human beings want to get an edge over other human beings by improving themselves, having better machine helpers. They’d also like to improve their children and have servants and so forth. Obviously, there’s a tremendous consumer demand. There’s also a military demand, which is worrisome for obvious reasons; namely, because the utility functions for military artificial intelligence are going to be things that could easily go awry.

Then there’s the drive towards exploration of space. Human bodies are very delicate; they are not radiation-hardened, they need water and supplies, and many things can wrong, as the Space Exploration program has shown. It would be much more efficient and inevitable to send cyborgs or artificial objects as the vanguards of space exploration. So, if we want to expand intelligence beyond the biosphere, that’s going to be an important drive. Let me draw some implications from these remarks, because they're meant also to stimulate discussion.

I don't think a singularity is imminent, although there has been quite a bit of talk about it. I don't think the prospect of artificial intelligence outstripping human intelligence is imminent because the engineering substrate just isn’t there, and I don't see the immediate prospects of getting there. I haven’t said much about quantum computing, other people will, but if you’re waiting for quantum computing to create a singularity, you’re misguided. That crossover, fortunately, will take decades, if not centuries.

There’s this tremendous drive for intelligence, but there will be a long period of coexistence in which there will be an ecology of intelligence. Humans will become enhanced in different ways and relatively trivial ways with smartphones and access to the Internet, but also the integration will become more intimate as time goes on. Younger people who interact with these devices from childhood will be cyborgs from the very beginning. They will think in different ways than current adults do.

Side by side with that, there will be autonomous intelligence and network intelligence. There will be a whole ecology of different kinds of powerful intelligence interacting with each other for decades. Now, that’s short on biological evolution timescales, but it’s reasonable on the timescale of human political and economic institutions. So, there will be the opportunity to evolve morality. That's a fortunate thing that there will be a possibility of learning by experience, interacting with different kinds of intelligence.

The idea that you can program morality, just like the idea that you can program other things that humans are good at, is very misguided. We just have to interact with the world and do them. That’s a big theme.

We’re very good at walking, at learning language, at constructing a three-dimensional world from partial information that arises in our retina, but we don’t know how we do any of those things. We learn to do them largely by interacting with the world. We understand even less how we learn morality or even what it is, but it comes from interacting with the world and other human beings. It’s fortunate that instead of a singularity there will be a time of coevolution, and that’s what the future of intelligence is going to look like.

* * * *

ROBERT AXELROD: I agree with your statement that AI and military use could easily go awry and, therefore, we need to be quite cautious about it. What about the analogy that autonomous vehicles could go awry? They’re already ten times better than humans.

WILCZEK: That reminds me of the talk we just heard about extreme cases. All you have to do is have a runaway vehicle that breaks down somehow.

AXELROD: Okay, in terms of accidents per mile driven, maybe ten is too much.

RODNEY BROOKS: That statistic is way off base.

AXELROD: It's not unreasonable to say that if they're not there now, they will be at least 1.2 times better than humans. In other words, an insurance company would rather insure an autonomous vehicle than a teenager.

BROOKS: This is the popular view in the press, and it is very misguided.

PETER GALISON: Because there's no data or because the data is the other way?

BROOKS: The data is much higher for cars, and the conditions under which they're driving is very different from how humans are driving. When you couple human pedestrians and human drivers, things change dramatically.

WILCZEK: This is a good example of the dangers of trying to solve complicated problems a priori without experience. We need practical experience with these things.

BROOKS: There has been a total turnaround in the automobile industry in the last three months on the predictions of when they're going to put cars out there. I'm actively involved in this area.

WILCZEK: The big message that I take from this analysis is that what's missing in artificial intelligence, and what humans do very well, is learn from the world. That's a very powerful source of information. If you can take information directly from the world by interacting, it may look slow, it may look inefficient, but the bandwidth of what's coming in is so enormous that it's worth it. Learning by doing should not be underestimated.

IAN MCEWAN: What is the physicist's view of the chances of making a self-conscious machine? Is there something in the nature of matter?

WILCZEK: Well, I can't speak for all physicists, but most physicists think that consciousness is an epiphenomenon. With all apologies, I don't think of it as a central problem.

SETH LLOYD: With due respect, consciousness is overrated. Ninety percent of the people I know are unconscious 90 percent of the time and the other 10 percent are unconscious 100 percent of the time.

There are different kinds of consciousness. There is the consciousness of running through the forest and not running into trees, which is a kind we'd like self-driving cars to have but they don't, and then there's a consciousness that I am a human being who is aware of myself as a human being—a kind of self-consciousness. This self-consciousness, which is the kind that humans often value and is the kind that they wonder whether machines can get, is what I'm talking about. Most human beings are unconscious by this definition.

JOHN BROCKMAN: Can an AI know what questions it should be asking? Or can it know what it doesn't know?

WILCZEK: Definitely, yes. Well, any question that you ask related to whether an AI can do something that humans are known to do, the answer is yes because it's overwhelmingly plausible that mind is based in matter. The human mind is based in matter and matter is what physics says it is.

There are certainly things about matter that we don't know, but for all practical engineering purposes, we know the fundamental laws as well as they're ever going to be known, and our knowledge is more than adequate to explain all observations. They've been tested in far more extreme conditions than you have in human brains, which are mild temperatures, mild densities, mild everything. Given that we know what matter is and that mind emerges from matter, we could in principle reproduce everything that goes on in a brain and nothing would be missing. I firmly believe, in that sense, natural intelligence is a special case of artificial intelligence. So, an engineered entity could do anything that a human can do.

BROCKMAN: So, when David Shaw was talking about his downloading brain capacity on a disc, you're saying you just replicate the brain.

WILCZEK: Replicate its function. That's very much a thought experiment. That's not practical at all. But as a matter of principle, it's hard to see how that could go wrong. People in physics do very delicate experiments and they have to correct for all kinds of possible sources of contamination.

NEIL GERSHENFELD: The mouse brain slicers are currently scanning brains down to the synaptic connection. We're just at the edge of having the first data sets that are good enough to do that from scratch. They have these crazy electron microscopes that have 100 beams that do nanometer slices that read every single synapse that they then reconstruct in 3D, so it's not that far off.

WILCZEK: That's very far from having a functional brain or segment of a brain.

CAROLINE JONES: Getting back to your strong and beautiful statement about engagement with the world and this human model of vast synaptic proliferation and then synaptic cropping that brings you down to an adult consciousness: How does the physicist's confidence in the material basis of consciousness jibe with this soft, meat machine creature that is in this environment? In other words, do you need to give Adam a skin made of sensory haptics? Do you need to have breath coming in and out of this machine to sense the world the way that you imagine the young human senses the world?

WILCZEK: No, I don't think it's necessary. Well, it depends what you want to do. Of course, if you want to make a human companion that humans get along with.

JONES: I would argue from the experience of art, and life, and feminist arguments that the meat is a big part of the epiphenomenon. No one imagines AI as needing meat, so that's part of my provocation. How much of the unconscious being in the world is part of the epiphenomenon that doesn't interest you but is hypothetically possible?

DAVID CHALMERS: There's a big industry working on artificial meat.

WILCZEK: Let’s talk about the embodiment of intelligence. If the idea that interacting with the world is a vital part of achieving general high levels of intelligence efficiently, then some kind of receptive apparatus is important. I don't think it would have to look like a human body, but it wouldn't be a bad idea to have a skin that’s telling you about the local environment.

Should you have two eyes as opposed to three, or four, or six? Does the skin need to be made out of flesh as opposed to some kind of plastic? These are very negotiable questions unless you want to have autonomous intelligences that interact intimately with humans. In that case, because humans are accustomed to interacting with other humans, it might be good to have the artificial guys look as human as possible.

Also, if you want to have artificial intelligence that appreciates the human experience and can make accurate models of what humans are thinking about and what they're experiencing, then again, you may want to have fairly accurate mimicry.

ALISON GOPNIK: I’m genuinely unsure about this, but it is striking that as long as we’ve had language and certainly as long as we’ve had writing, it’s a real question about how abstract we can get intimate, close, social interactions. Think about Elizabeth Barrett and Robert Browning, right? It’s remarkable that with the technology of having a quill and a piece of paper, you can have a completely different medium that doesn’t look like a typical human medium of interaction at all. You seem to be able to get all the complexities, the interactions, and all the subtleties working just fine.

WILCZEK: They had a pretty good model of what they were dealing with.

BROOKS: Someone talked about the uncanny valley. I had tried with my graduate students when I was at MIT and I tried with people in my company to build a three-armed robot because we can optimize much more. I’ve never been able to get anyone to build a three-armed robot. They feel it’s too icky. They won't do it.

WILCZEK: Why don’t you have them do eight and call it an octopus?

BROOKS: I'm not saying they're right, I'm just pointing out that there's a barrier people have, which is not bounded rationality or anything like that. No matter what a robot looks like, how it looks is making a promise of what it's going to deliver. And when that promise isn't matched by what it does deliver, it's really upsetting.

GOPNIK: Studies have just come out that kids do not experience an uncanny valley in the same way that adults do. You would have thought that that's the natural state and then we have to overcome it, but it may be a result of a whole lot of experience with machines. Maybe this is a generational effect.

WILCZEK: You called it a "barrier," which may be appropriate. A barrier is something you can reach or get over, and once you've reached it maybe there's a smooth path after that—acceptance. A large part of the unease is simply not knowing what to expect.

JONES: The classic formation of the uncanny valley is if it gets too close to the human, it's profoundly disturbing. It's experienced as a creepy freak, right? The eight arms would be the way to go and the twenty-eight eyes would be the way to go.

LLOYD: Frank, is your main point that this is going to happen, but it's going to happen slowly?

WILCZEK: Yes.

LLOYD: We're going to have time to get used to this, so maybe we should practice being nice to these artificial intelligences before we let them appoint our president.

WILCZEK: Yes. A feeling of humility and learning by doing, not only practical tasks but also the coevolution of the different kinds of intelligences, will be something that evolves and involves learning by doing.

BROOKS: In the relatively short term, and by short term I mean the next ten to twenty years, as we get more robots in our homes, largely driven economically by the need for elder care, those robots are going to have very

different umwelts than humans. They’re going to have all the senses that are cheap as anything because they aren’t here. They’re going to be able to detect any Bluetooth device, any Wi-Fi device, they’re going to use Wi-Fi to be able to detect when someone is breathing, they’re going to be able to see the hotspots where someone was just sitting on the couch.

You have to have some intelligence to take care of it. But they’re going to have a very different sensory perception of the world that they’re sharing with us. How we get used to them will be interesting. Will there be certain species of robots with particular sorts of sensory stuff that we understand, or will we be continuously surprised by them knowing stuff about us that we didn’t expect them to know?

W. DANIEL HILLIS: Don’t you think we’ll have very different sets of perception by then, too?

BROOKS: Well, yes. Maybe not in that ten to twenty-year timeframe.

WILCZEK: There are two ways of talking about AI that are very common that are not appropriate, and it’s going to become increasingly clear that they’re not appropriate. One is to talk about AI in terms of "us versus them." They're our creations and we will be interacting in very intimate ways with them. They’ll be part of society.

The other thing that you were alluding to is that it’s common to talk of AI as if it’s one thing. Intelligence, whether natural or artificial, can take many forms. Natural intelligence is embodied in all kinds of animals, and maybe even in our digestive system and immune system. Artificial intelligence is all kinds at all levels. Some people would argue that thermostats are a form of artificial intelligence. Then you have distributed intelligence, you could have soldiers, you could have servants, and those would be very different kinds of minds.

NEIL GERSHENFELD
Morphogenesis for the Design of Design

As we work on the self-reproducing assembler, and writing software that looks like hardware that respects geometry, they meet in morphogenesis. This is the thing I'm most excited about right now: the design of design. Your genome doesn't store anywhere that you have five fingers. It stores a developmental program, and when you run it, you get five fingers. It's one of the oldest parts of the genome. Hox genes are an example. It's essentially the only part of the genome where the spatial order matters. It gets read off as a program, and the program never represents the physical thing it's constructing. The morphogenes are a program that specifies morphogens that do things like climb gradients and symmetry break; it never represents the thing it's constructing, but the morphogens then following the morphogenes give rise to you.

What's going on in morphogenesis, in part, is compression. A billion bases can specify a trillion cells, but the more interesting thing that's going on is almost anything you perturb in the genome is either inconsequential or fatal. The morphogenes are a curated search space where rearranging them is interesting—you go from gills to wings to flippers. The heart of success in machine learning, however you represent it, is function representation. The real progress in machine learning is learning representation. How you search hasn't changed all that much, but how you represent search has. These morphogenes are a beautiful way to represent design. Technology today doesn't do it. Technology today generally doesn't distinguish genotype and phenotype in the sense that you explicitly represent what you're designing. In morphogenesis, you never represent the thing you're designing; it's done in a beautifully abstract way. For these self-reproducing assemblers, what we're building is morphogenesis for the design of design. Rather than a combinatorial search over billions of degrees of freedom, you search over these developmental programs. This is one of the core research questions we're looking at.

NEIL GERSHENFELD is the director of MIT's Center for Bits and Atoms; founder of the global fab lab network; the author of *FAB*; and co-author (with Alan Gershenfeld & Joel Cutcher-Gershenfeld) of *Designing Reality.*

* * * *

NEIL GERSHENFELD: I'd like to end this interesting long day by explaining why I think computer science was one of the worst things ever to happen to computers or science, why I believe that, and what that leads me to. I believe that because it's fundamentally unphysical. It's based on maintaining a fiction that digital isn't physical and happens in a disconnected virtual world.

One of my students built and runs all the computers Facebook runs on, one of my students used to run all the computers Twitter runs on—this is because I taught them to not believe in computer science. In other words, their job is to take billions of dollars, hundreds of megawatts, and tons of mass, and make information while also not believing that the digital is abstracted from the physical. Some of the other things that have come out from this lineage were the first quantum computations, or microfluidic computing, or part of creating some of the first minimal cells.

Stephen made the observation that we're surrounded by computation, most of which we don't use. This is what leads me to wanting a do-over. I view the current state of computer science as a bit like *Metropolis*, where it's training people to frolic in the garden while somebody in the basement moves the levers. What I want to talk about is how you bring them together.

First of all, I've come to the conclusion that this is a historical accident. I could ask Marvin what John von Neumann was thinking, and I could ask Andy Gleason what Turing was thinking, and neither of them intended us to be living in these channels. Von Neumann wrote beautifully about many things, but computer architecture wasn't one of them. We've been living with the legacy of the EDVAC and the machines around us, and much of the work of computers is not computationally useful because it's just shuttling stuff. The Turing machine was never meant to be an architecture. In fact, I'd argue it has a very fundamental mistake, which is that the head is distinct from the tape. And the notion that the head is distinct from the tape—meaning, persistence of tape is different from interaction—has persisted. The computer in front of Rod Brooks here is spending about half of its work just shuttling from the tape to the head and back again.

There’s a whole parallel history of computing, from Maxwell to Boltzmann to Szilard to Landauer to Bennett, where you represent computation with physical resources. You don’t pretend digital is separate from physical. Computation has physical resources. It has all sorts of opportunities, and getting that wrong leads to a number of false dichotomies that I want to talk through now. One false dichotomy is that in computer science you’re taught many different models of computation and adherence, and there’s a whole taxonomy of them. In physics there’s only one model of computation: A patch of space occupies space, it takes time to transit, it stores state, and states interact—that’s what the universe does. Anything other than that model of computation is physics and you need epicycles to maintain the fiction, and in many ways that fiction is now breaking.

I’ve been working with people on exascale computer architecture, the biggest super computer architecture. If you look at what it costs to move data to memory, what it costs to do interconnect, and what it costs to have all the processors working usefully, all of those things are breaking. We did a study for DARPA of what would happen if you rewrote from scratch a computer software and hardware so that you represented space and time physically. So, if you zoom from a transistor up to an application, you change representations—completely unrelated ones—about five different times. If you zoom the building we’re in from city, state, country, it’s hierarchical, but you respect the geometry. It turns out you can do that to make computer architectures where software and hardware are aligned and not in disconnected worlds. One of the places that I’ve been involved in pushing that is in exascale high-performance computing architecture, really just a fundamental do-over to make software look like hardware and not to be in an abstracted world.

Right now, we’re in deep-learning mania as one of the things pushing computing. Depending on how you count, this is now the fifth boom-bust cycle. From a distance it looks like we’re now in a boom cycle. This is the good thing. A quiet trend that’s been emerging is that scaling is what's driving the current AI boom—networks gathering more data, bigger memories storing the data, more processing cycles. It's a quiet, really interesting trend as it turns out. Most of what’s getting the attention on the deep-learning architectures don’t matter much. Many different approaches work equally well. There’s nothing magic about the deep-learning

architectures. The magic is there's more data with more memory with more cycles. It's a cargo cult, the obsession with the acronym zoo of deep learning. It's just an exercise in scaling that's been making that possible.

Analog versus digital are not two distinct choices where you can pick one or the other. What's interesting is what lies between them. As an example, my lab spun off a chip company that uses analog degrees of freedom to solve digital problems. A digital system lives on the corner of a hypercube, but what we did in that chip company was use the analog device degrees of freedom to go through the interior of the hypercube, not to stay on the corners. It saves power and speeds and has all these performance benefits.

That's not a new idea in the context of optimization. Like many of the largest scale computations, what's used are things called interior point methods, or relaxations, where you have a discrete answer you want—like routing an airplane or which way to turn a car—but the way you get through it is to relax the discrete constraints and use internal degrees of freedom. These interior point methods are the most important algorithms for solving large-scale computational problems. If you just took one of my chips doing a physical version of this, a neurobiologist would have absolutely no idea what was going on in it, but it would make perfect sense in an introductory optimization class.

Digital isn't ones and zeroes. One of the hearts of what Shannon did is threshold theorems. A threshold theorem says I can talk to you as a wave form or as a symbol. If I talk to you as a symbol, if the noise is above a threshold, you're guaranteed to decode it wrong; if the noise is below a threshold, for a linear increase in the physical resources representing the symbol there's an exponential reduction in the fidelity to decode it. That exponential scaling means unreliable devices can operate reliably.

The real meaning of digital is that scaling property. But the scaling property isn't one and zero; it's the states in the system. In the end, what these interior point and relaxation methods do is drive to an outcome that's a discrete state, but you pass through continuous degrees of freedom. It's very naïve to say digital is ones and zeroes. It's state restoration, but you can use continuous degrees of freedom. In many different areas this is done to do the state restoration.

Now: threshold theorems. It was first proved by Shannon. Von Neumann applied Shannon to computing to show how reliable computers can operate with unreliable devices, but the thing that excites me is threshold theorems were invented four billion years ago, which is the evolutionary age of the ribosome. The connection there is if you mix chemicals and make a chemical reaction, a yield of a part per 100 is good. When the ribosome—the molecular assembler that makes your proteins—elongates, it makes an error of one in 10^{4}. When DNA replicates, it adds one extra error-correction step, and that makes an error in 10^{-8}, and that's exactly the scaling of threshold theorem. The exponential complexity that makes you possible is by error detection and correction in your construction. It's everything Shannon and von Neumann taught us about codes and reconstruction, but it's now doing it in physical systems.

One of the projects I'm working on in my lab that I'm most excited about is making an assembler that can assemble assemblers from the parts that it's assembling—a self-reproducing machine. What it's based on is us. We're made from 20 parts, amino acids, and what's interesting about amino acids is they're not interesting. They have simple properties like hydrophobic and hydrophilic and basic and acidic, but you can compose them to make muscles and motors and sensors. In the same way, we're taking 20 inorganic properties like conducting and insulating to show you can compose them hierarchically. In fact, the project funding was a proposal to the DoD to reduce their whole supply chain to 20 parts, these fundamental building blocks, and they're based on digitizing the materials.

Compare state of the art manufacturing with a Lego brick or a ribosome: When a kid plays with Lego, you don't need a ruler because the metrology comes from the parts. It's the same thing for the amino acids. The Lego tower is more accurate than the motor control of the child because you detect and correct errors in their construction. It's the same thing with the amino acid. There's no trash with Lego because there's information in the construction that lets you deconstruct it and use it again. It's the same thing with the amino acids. It's everything we understand as digital, but now the digital is in the construction. It's digitizing the materials. The core project of assembling an assembler is, in part, a paradigmatic challenge. If you look at scaling coding construction by assembly, ribosomes are slow—they run at one hertz, one amino acid a second—but a cell can have a

million, and you can have a trillion cells. As you were sitting here listening, you're placing 10^{18} parts a second, and it's because you can ring up this capacity of assembling assemblers. The heart of the project is the exponential scaling of self-reproducing assemblers.

As we work on the self-reproducing assembler, and writing software that looks like hardware that respects geometry, they meet in morphogenesis. This is the thing I'm most excited about right now: the design of design. Your genome doesn't store anywhere that you have five fingers. It stores a developmental program, and when you run it, you get five fingers. It's one of the oldest parts of the genome. Hox genes are an example. It's essentially the only part of the genome where the spatial order matters. It gets read off as a program, and the program never represents the physical thing it's constructing. The morphogenes are a program that specifies morphogens that do things like climb gradients and symmetry break; it never represents the thing it's constructing, but the morphogens then following the morphogenes give rise to you.

What's going on in morphogenesis, in part, is compression. A billion bases can specify a trillion cells, but the more interesting thing that's going on is almost anything you perturb in the genome is either inconsequential or fatal. The morphogenes are a curated search space where rearranging them is interesting—you go from gills to wings to flippers. The heart of success in machine learning, however you represent it, is function representation. The real progress in machine learning is learning representation. How you search hasn't changed all that much, but how you represent search has. These morphogenes are a beautiful way to represent design. Technology today doesn't do it. Technology today generally doesn't distinguish genotype and phenotype in the sense that you explicitly represent what you're designing. In morphogenesis, you never represent the thing you're designing; it's done in a beautifully abstract way. For these self-reproducing assemblers, what we're building is morphogenesis for the design of design. Rather than a combinatorial search over billions of degrees of freedom, you search over these developmental programs. This is one of the core research questions we're looking at.

I started off this diatribe by complaining about computer science, but von Neumann and Turing ended exactly here. The last thing von Neumann

worked on—and this is something he wrote beautifully about—was self-reproducing machines. If you've ever read it, his memo on the EDVAC is a mess. The programming manual where the von Neumann architecture emerged is a dreadful document. It's a mess. What he wrote about self-reproducing machines was exquisite. It's a beautiful posthumous document asking how a thing can communicate a computation for its construction, how to abstract a self-reproducing thing. He was asking it to get at the heart of what is life. It was a theoretical thing at that time. That's what he ended his life doing. The last thing Turing ended his life doing was studying morphogenesis. What it's casually known for is Turing spots and patterns, but that was the detail. What he was really asking was bits from atoms or atoms from bits. He was asking, how do genes give rise to us?

Looking at exactly this question of how a code and a gene give rise to form. Turing and von Neumann both completely understood that the interesting place in computation is how computation becomes physical, how it becomes embodied and how you represent it. That's where they both ended their life. That's neglected in the canon of computing, but we're now at this interesting point where I'm on the hook to deliver on a research program to make a self-reproducing von Neumann assembler. We can think about making these things now, of embodying it. It is a third digital revolution. There is communication, then computation, now fabrication. It's not a separate one, but it merges them because it merges them in a thing that communicates its construction to fabricate.

At MIT the first real-time computer was the Whirlwind. Then came the PDP as the mini computers, and there were thousands of those. Then came the hobbyist computers like the Altair, and there are millions of those. Then came the personal computers and smart phones, and there are billions of those. Now, there are the Internet of Things devices, and there are trillions of those.

The Nest thermostat, roughly, has the capacity of the PDP—computing scale from one to a thousand to a million to a billion to a trillion. You could see all of that lurking in 1965 when Gordon Moore made his first plot of Moore's law that scaled for fifty years. In the same way, if you take digital fabrication, it's been scaling for about a decade in the same way. You can make a

Moore’s law-like plot for performance and scaling of digital fabrication, and there’s a close historical parallel.

MIT made the first NC mill in 1952. That’s like the mainframe. For NSF I started setting up FAB labs, which are mini versions of the big lab I run. With current digital fab tools, they would fit in a room like this—and that’s like the PDP version. There’s a thousand of those today. We’re using those to make machines that make machines, not self-reproducing assemblers but rapid prototyping tools that make rapid prototyping tools, and that’s moving towards a million of them.

In the lab, we’re developing these assemblers that I described and then working toward the self-assemblers. All those things exist in some form today, but they’re going to be emerging between now and fifty years from now, but you can see the thousand, million, billion, trillion scaling happening for digital fabrication.

We're at an interesting point now where it makes as much sense to take seriously that scaling as it did to take Moore’s law scaling in 1965 when he made his first graph. We started doing these FAB labs just as outreach for NSF, and then they went viral, and they let ordinary people go from consumers to producers. It’s leading to very fundamental things about what is work, what is money, what is an economy, what is consumption.

There’s legislation in the Senate and House right now for universal access to digital fabrication, like there was for communication and computation. We're also working with Bhutan’s prime minister—the country is based on gross national happiness, but they buy crap trucked in from India—on how to make gross national happiness physical.

We’re working with a number of cities around the world that have failed economies on how to turn consumption into creation. In the same way that the Internet emerged in the mini computer era, this fifty-year scaling of digital fabrication is emerging today, and the equivalent of "how does the Internet work?" is growing up around it. A surprising fraction of my time has just gone into working with all these governments and organizations and social groups on if anybody can make anything anywhere, how does that reinvent societies and economies?

I started with complaining that computer science was the worst thing to happen to computers or science because it’s unphysical, and pointed out that you can have a do-over of computer science that’s much more aligned with physics. It has all kinds of benefits ranging from computing with very different physical systems to limits of high-performance computing but, ultimately, reuniting computer science and physical science leads to merging the bits and atoms. Fabrication merges with communication and computation. Most fundamentally, it leads to things like morphogenesis and self-reproducing an assembler. Most practically, it leads to almost anybody can make almost anything, which is one of the most disruptive things I know happening right now. Think about this range I talked about as for computing the thousand, million, billion, trillion now happening for the physical world, it's all here today but coming out on many different link scales.

The last time we gathered, there was a suggestion to turn it into a book, which was a lovely exercise. Coming here, John asked me what I thought we should do coming out from this. I had three suggestions that he thought were all horrible, so I’ll end with those. The baseline is we have a lovely weekend, we admire each other, and then we go home. So that’s the default.

One suggestion I have comes from a conversation with my younger brother who led the biggest video game studio, Activision, and he was horrified when he discovered when you write a book it’s good if it sells thousands of copies. He’s used to selling tens of millions of whatever he does. He left Activision. He now has a company that does games for education and social change. The most recent one they did that got a lot of attention was with Alaska native storytellers. There are great traditions, but terrible alcoholism, and suicide, and unemployment, and they worked with Alaska native storytellers to tell narratives in immersive video game experiences. There’s a whole bunch of examples like that. One suggestion I had John hated was we build the world we’re describing as an immersive experience and get it in the hands of millions of people.

I had done a number of friend-of-friend movie advising in Hollywood, and that led through a collaboration where I helped start an office called the Science Entertainment Exchange, which hijacks popular media. It takes movies and TV shows and uses them as covers to put in science teaching,

and it's been working really well embedding science in all kinds of popular shows. The second idea I had was we take everything we're trying to do and embed it in the popular conscious by hijacking some movies or TV shows.

Then the third one has been working with some interesting groups that put together bit stadium shows, and so this has been lovely, but it's just for us. We do this on an epic scale was a third suggestion. Those are the three ideas John thought were terrible that I'll conclude with, so now I'll step back and open that for discussion.

* * * *

PETER GALISON: I wonder if we could discuss whether there's something different about the biological case than, say, the physical properties that lead to snowflakes or crystals. That is to say, elementary atomic forces don't have encoded this complicated hexagonal form, but you get there. They just make local decisions, and the local decisions add up like Legos to something else. My question is about the physicalization or the embodiment of computation.

I can think of several reasons why you might want to shorten or eliminate the gap between software and hardware. One might be that there's an aesthetic objection that something's wrong with hardware that is disjunct from the way we represent it. There are other things that we do in our representation where they're not matched; for example, differential equations don't look like the things that they're often representing. Or another might be efficiency, that if we could somehow have software that matched the physicality of, say, atoms and bits, it would run without the frictional loss of computing power in our everyday devices. Another might be that there are more and more cases where the software is embedded in the hardware itself. If you dig into your Intel chip, there's a lot of software in them before you get to high-level programming. Suppose we agree that there is this gap between the representation and the things represented, what is it that propels you?

GERSHENFELD: For your first passing point about the snowflake, I'll make a passing point. The work I'm describing on coding assembly of digital materials isn't a single-length scale. We're doing that in molecular biology

when we make synthetic cells. We're doing nanofab to make nanostructures. We're micromachining microstructures up to where we're working with Airbus and robots to make jumbo jets and NASA to make spaceships on big scales. What I spoke about isn't a single-length scale. It's better to think about it as the dynamic range between the smallest feature you need to control and the size of the system.

Why align computer science and physical science? There are at least five reasons for me. Only lightly is it philosophical. It's the cracks in the matrix. The matrix is cracking. 1) The fact that whoever has their laptop open is spending about half of its resources shuttling information from memory transistors to processor transistors even though the memory transistors have the same computational power as the processor transistors is a bad legacy of the EDVAC. It's a bit annoying for the computer, but when you get to things like an exascale supercomputer, it breaks. You just can't maintain the fiction as you push the scaling. The resource in very largescale computing is maintaining the fiction so the programmers can pretend it's not true is getting just so painful you need to redo it. In fact, if you look down in the trenches, things like emerging ways to do very largescale GPU program are beginning to inch in that direction. So, it's breaking in performance.

2) We're just wasting resources. When you look at what's going on in your Intel chip, it's right at the edge of analog. They do a lot of work. Inside it's awfully analog but ends up looking digital on the outside. We're wasting a lot of the computational power of the transistor. With the chip fab I mentioned, we're wasting degrees of freedom in the devices that aren't a simple version of analog versus digital. You can solve digital problems, but by using the analog degrees of freedom, you win speed, power, performance, and all kinds of good stuff.

3) When we were in the early days of quantum computing or the stuff we did on microfluidic logic, you're computing with fundamentally different physical resources where you need to represent the computation in a way that can describe the physics that you're working with.

4) The final reason goes back to where von Neumann ended up. When I make this self-reproducing assembler in the very short term, I'm using conventional computer architectures for the intelligence of it, but what I

need to do is overlay the computation as geometry. If I'm doing morphogenesis with a self-reproducing system, I don't want to then just paste in some lines of code. The computation is part of the construction of the object. I need to represent the computation in the construction, so it forces you to be able to overlay geometry with construction.

There are all different reasons, but they all lead you to the same place. Interestingly, for the do-over I mentioned in DARPA, we took the BLAS, which are the routines that underlie high-performance computing, and we rewrote them in a geometrical spatial computing model. What's interesting is a lot of the things that are hard—for example, in parallelization and synchronization—come for free. By representing time and space explicitly, you don't need to do the annoying things like thread synchronization and all the stuff that goes into parallel programming.

DAVID CHALMERS: What you're saying is, when thinking about software, hardware and physics matter. In some sense everyone has known all along that the hardware matters and the physics matters, and chip makers and everyone else under the sun has been thinking about how to do the best computing you can given the limitations that you have about technology and the resources of physics. One thing you're saying is that we haven't done everything we can to take advantage of the hardware possibilities, so we've got to push the project harder and faster. Does it go beyond that? The part about fabrication and self-assembly is fundamentally new and different.

GERSHENFELD: Let me help you connect those parts. Communication degraded with distance. Along came Shannon. We now have the Internet. Computation degraded with time. The last great analog computer work was Vannevar Bush's differential analyzer. One of the students working on it was Shannon. He was so annoyed that he invented our modern digital notions in his Master's thesis to get over the experience of working on the differential analyzer.

Today, in this computer, it's head-bangingly stupid what's going on with this accidental legacy of von Neumann architecture. He never talked about the von Neumann architecture long past its due date. Much of the resources are shuttling information from memory transistors to processor transistors,

wasting the power of all of this, and then the utilization of it is more inefficient still when you go from the software compilation to the hardware.

So, one of the points was just it's very inefficient. It doesn't matter if you're doing word processing; it does matter if you're pushing limits of computing performance. So, very low power or very high power, you care about that.

CHALMERS: If it's so head-bangingly stupid, why didn't someone from Intel figure this out years ago?

GERSHENFELD: What's interesting is that there's a whole parallel history. We've been lulled into sleep by Gordon Moore. I spent some time with Gordon Moore in the early days of this fabrication scaling I was mentioning, and he was amused by the parallel with what he did at that time. It's like the matrix. We had a few decades where we could pretend that nobody's moving the levers in the basement and we can frolic in the garden. There's been a parallel history all the way through it. It passes through people like Danny. There are a number of device physics. There's a whole parallel history building this, but you could ignore it. Again, limits of either high performance, low power are pushing it.

I started by mentioning my students who built the computers for Facebook and Twitter, and they're not doing this at the fundamental physics level, but they had to completely re-architecture how you build a data center with coarse-grain versions of it. You don't see it, but it percolates in things like how Jason built the Facebook data center. Just to recap the answer, you need to do what I'm describing if you don't compute with Intel. So, the stuff we did on quantum computing or fluidic or molecular computing, you need to revisit these assumptions.

If you are confused by everything I say, and you take a single thing away, it's the last part I talked about, about digitizing fabrication. It's not about computing and then there's this other thing here, but it's the synthesis. When you merge communication with computation with fabrication, it's not there's a duopoly of communication and computation and then over here is manufacturing; they all belong together. The heart of how we work is this trinity of communication plus computation and fabrication, and for me the real point is merging them.

W. DANIEL HILLIS: I was going to give just a very specific small example that supports the abstraction that you're saying. In modern ways of analyzing algorithms, and computers, and the computer science, we count the cost of moving a bit in time. We call that storage, and that's very carefully measured in the algorithms and things like that. The cost of moving a bit in space is completely invisible, and it just doesn't come up. There's no measure of that in the way that we abstract it, but if you look at the megawatts that are dissipated in high-performance computers, it mostly comes from moving bits in space. That's the big limitation, and that's also where the errors are and where the cost is. So, our abstraction that we're thinking about the algorithms in is completely out of sync with where our costs are.

CHALMERS: You mean that hardware makers have not been thinking about those costs of moving bits in space?

GERSHENFELD: One more example of the cracks in the matrix is, every few months there's a headline about a new security vulnerability, and an awful lot of them have to do with things that are supposed to be far away in computation space colliding in physical space, because there's no way to say things that are far apart computationally should be far apart physically.

I've spent time with the people after Gordon who ran Moore's law at Intel, the keepers of Moore's law, and one of the most evocative images came from one of them describing his job as the scene in the Indiana Jones movie when the boulder is running down. All he can do is not get run over by the boulder. They're running this multibillion-dollar oil tanker, and it's hard to steer. They have to make sure the boulder doesn't run over them.

I almost took over running research at Intel. It ended up being a bad idea on both sides, but when I was talking to them about it, I was warned off. It was like the godfather: "You can do that other stuff, but don't you dare mess with the mainline architecture." We weren't allowed to even think about that. In defense of them, it's billions and billions of dollars investment. It was a good multi-decade reign. They just weren't able to do it.

SETH LLOYD: Maybe we take what Frank and others have been saying about the power of the brain and ask what we would need in a

computational device to do that. The brain has, we were saying, 10_{11} neurons, around 10_{15} connections, and it operates at the 100-hertz scale. Suppose you wanted to get a silicon device that had similar scale. If the size of the objects were a nanometer and you weren't worrying about the wiring, you would have to have about one electron per transistor. You'd have to go down to single-electron transistors. This device would be tremendously noisy, the problem of moving information around. If you want to get to the kind of information processing that human beings and other animals have, you would need to go far beyond the paradigms that people have: dealing with noisy computation, making it analog, mapping the way the physical processing is going on onto a chip in a way that's very different from the way that people do the architecture right now, doing things massively in parallel.

If you wish to fulfill the promise of Moore's law to get artificial intelligences that are similar in scale to human beings, you've got to do something quite different.

GERSHENFELD: Analog doesn't mean analog. In other words, analog in this context means you have states, and you recover from errors, and you detect states. But states are outcomes of the system, they're not ones and zeroes. One of the things we're stuck in is this idea that a state is one and a zero. This device in front of me keeps recurring the state not at the high-level thing I'm trying to do, but at the ones and zeroes.

These interior point relaxation methods I was describing both in software optimization and in emerging chips do digitize, but they're digitizing on high-level outcomes but using the analog degrees of freedom. That was behind my comment that when the brain does a few moves a second, it's moving through this very high-dimensional space, ending into a discrete outcome. So, the effective number of operations that are done this way is an enormous number.

TOM GRIFFITHS: I wanted to return to Rod's talk, asking whether any of the things you learn about when you're thinking about scaling should inform the way that we think about neuroscience in terms of getting at some of the inadequacies of classic models of computation for neuroscientists.

GERSHENFELD: I was recently at a retreat of many of the leading neuroscientists for a review of the state of the art of the field, and boy I was horrified. They were horrified. The state of the art of neuroscience is like you throw the watch at the wall and you see the parts that come out. We had a lively discussion about the devices I'm building and the algorithms we're using they would be completely stumped by. They would have absolutely no idea how to recognize that was going on. We don't have an easy next step after that, but there's an interesting dialogue with the neuroscientists about it.

ALISON GOPNIK: There is something that's a bit puzzling about this, which is that you have these incredibly complex devices—brains—and they can be translated into a bunch of symbols on a piece of paper or a bunch of simple digitally described symbols in a language, and that seems to be able to do a lot of work for human beings. Arguably, a lot of the capacities for intelligence that we have come from things like being able to talk to one another, or write, or use symbols in these ways that from a hardware perspective are completely trivial.

I'm not being disingenuous about this. This is a real puzzle, and in some ways what Turing is modeling, what he's starting out with when he's thinking about the computer who's sitting there in Bletchley Park is not anything like this tiny bit of complexity compared to the complexity of what's going on underneath the hood. It's puzzling to me about what the relationship is between those two things.

GERSHENFELD: One interesting group I worked with was at Wright-Patterson Air Force Base where among the most sensory overloaded tasks are fighter pilots, and so they wanted to make planes you could fly by thinking. What came out of that after a lot of work is that it's a terrible idea. The reason is, with a lot of work to pull a lot of signals out and do a lot of interpretation, you can barely control anything because all of this just isn't the right representation. All of this is designed so that this moves and this moves, and the best way to interface with this is to move your fingers. So, this representation is an internal one and then this is an external one.

GOPNIK: It seems to me like it's an incredibly interesting understudied fact that what this all ends up driving is a bunch of fingers and your larynx. This

tiny system with tiny degrees of freedom and very little complexity is the thing that's doing the work that we think of as being a lot of the work of intelligence.

GERSHENFELD: But again, these kinds of relaxation interior point methods that I keep alluding to, there's something similar to them in that they're moving through these billion-dimensional spaces, but what they're putting outside is not the interior point but statistics of the states that they're getting driven to. So, there are analogs between unpacking the huge number of internal degrees of freedom versus small numbers of observable degrees of freedom in these engineered systems.

CHALMERS: The brain also has these amazing hardware inefficiencies in it, which are analogous to your hardware cases, like the fact that it uses electrical transmission within neurons, but between cells it's chemical transmission. So, I guess the brain just got locked into that the way Intel got locked in years ago, and then it couldn't escape the boulder fast enough.

GERSHENFELD: That's true. Again, the embodiment of everything we're talking about, for me, is the morphogenes—the way evolution searches for design by coding for construction. And they're the oldest part of the genome. They were invented a very long time ago and nobody has messed with them since.

LLOYD: I disagree with that about the brain. The electrical signals use a lot more power, but they go fast and they go a long distance. The synaptic connections, of which there are thousands more, use much less power. I'm talking about just energy, but they go over a very tiny distance and they only use a few hundred molecules. So, it's pretty efficient.

CAROLINE JONES: They're chemical, and there's a kind of redundancy and robustness in those separate things. There's also a different system of feedback, which is fascinating. The chemicals are regulated by completely different body systems, which allows for all different kinds of intelligence to overlap and reinforce each other.

GOPNIK: It's worth pointing out that plasticity is expensive. This is one of my favorite factoids: Everyone knows brains are taking about 20 percent of

calories. If you look at four-year-olds, it's 66 to 70 percent of calories are getting used up by brains. It's not so much that they're doing the computations, but they're establishing what the wiring looks like.

GERSHENFELD: I worked with an IBM largescale computer architect on a project to make a computer that can physically remodify itself—taking the kind of assembler I'm describing to make a computer that can rebuild its construction. We're still discussing that and working on it, but he told me something interesting. They did an early crude version of that, and what they discovered was the computer got configured but never reconfigured, which is very analogous to learning. The configurability was used to adapt the computer to the workload, but they never went back to change it. So, that led us to look at not reconfigurable but just configurable computers, like computers that can build themselves but don't necessarily need to unbuild themselves.

Get over digital and physical are separate; they can be united. Get over analog as separate from digital; there's a really profound place in between. We're at the beginning of fifty years of Moore's law but for the physical world. We didn't talk much about it, but it has the biggest impact of anything I know if anybody can make anything.

I'll leave you with my three questions that John doesn't like. Do you want to make a video game for millions of people to live in the world we're in? By the way, I did one of these. It's fun to build the world you're trying to create. Do you want to portray it on a large scale? Do you want to do what we're doing here on a large scale? Any of those have great teams that could help with it rather than just doing a book next.

DAVID CHALMERS
The Language of Mind

Will every possible intelligent system somehow experience itself or model itself as having a mind? Is the language of mind going to be inevitable in an AI system that has some kind of model of itself? If you've just got an AI system that's modeling the world and not bringing itself into the equation, then it may need the language of mind to talk about other people if it wants to model them and model itself from the third-person perspective. If we're working towards artificial general intelligence, it's natural to have AIs with models of themselves, particularly with introspective self-models, where they can know what's going on in some sense from the first-person perspective.

Say you do something that negatively affects an AI, something that in an ordinary human would correspond to damage and pain. Your AI is going to say, "Please don't do that. That's very bad." Introspectively, it's a model that recognizes someone has caused one of those states it calls pain. Is it going to be an inevitable consequence of introspective self-models in AI that they start to model themselves as having something like consciousness? My own suspicion is that there's something about the mechanisms of self-modeling and introspection that are going to naturally lead to these intuitions, where an AI will model itself as being conscious. The next step is whether an AI of this kind is going to naturally experience consciousness as somehow puzzling, as something that potentially is hard to square with basic underlying mechanisms and hard to explain.

DAVID CHALMERS is University Professor of Philosophy and Neural Science and co-director of the Center for Mind, Brain, and Consciousness at New York University. He is best known for his work on consciousness, including his formulation of the "hard problem" of consciousness.

* * * *

DAVID CHALMERS: John brought us together to talk about possible minds—minds in human and AI systems and the variety of minds, not just that there are but that could be. I think about the mind for a living,

especially the human mind. The mind is something that we all know we have. When it comes to AI systems, AI researchers are not quite sure what to make of this. All sorts of questions arise: What is it? What would it be for an AI system to have a mind? What's the research project?

Today, I'm just going to talk about an angle on thinking about the mind and the mind-body problem that also suggests a research program in AI that might help us bite off a little bit of the big philosophical puzzles around the mind and its relationship to the brain.

We've got these bodies and these brains, which work okay, but we also have minds. We see, we hear, we think, we feel, we plan, we act, we do; we're conscious. Viewed from the outside, you see a reasonably finely tuned mechanism. From the inside, we all experience ourselves as having a mind, as feeling, thinking, experiencing, being, which is pretty central to our conception of ourselves. It also raises any number of philosophical and scientific problems. When it comes to explaining the objective stuff from the outside—the behavior and so on—you put together some neural and computational mechanisms, and we have a paradigm for explaining those.

When it comes to explaining the mind, particularly the conscious aspects of the mind, it looks like the standard paradigm of putting together mechanisms and explaining things like the objective processes of behavior leaves an explanatory gap. How does all that processing give you a subjective experience, and why does it feel like something from the inside doesn't look like it's directly addressed by these methods? That's what people call the hard problem of consciousness, as opposed to, say, the easy problems of explaining behavior.

Discussion can then spin off in a thousand directions. Could you explain conscious experience in terms of the brain? Does it require something fundamentally new? Does it exist at all? Lately, I've been interested in coming at this from a slightly different direction. We've got the first-order problem of consciousness, and then it's often hard for people from AI research, or neuroscience, or psychology to say, "There's a problem here, but I'm not quite sure what I can do with it."

The angle I've been thinking about lately is to step up a level. I don't know where this slogan comes from, "Anything you can do, I can do meta." Sometimes it's attributed to my thesis advisor, Doug Hofstadter, but I don't think it was him. I've seen it attributed to Rudolf Carnap, but I don't think it was him, either. In any case, I've lately been thinking about what I call the *meta*-problem of consciousness. The first-order problem of consciousness explores how all this processing gave rise to a conscious experience. The meta-problem asks why we think there is a problem of consciousness and, in particular, why we go around saying there is a problem of consciousness.

Belief in consciousness and belief in the problems of consciousness is extremely widespread. So, it's consistent with this approach, by the way, that it will all be an illusion or nonsense. Nonetheless, there's an interesting psychological problem. It is a fact of human behavior that people go around saying things like, "Hey, I'm conscious." They go around reporting subjective experience. Even in kids you can get various puzzlements that you would associate with conscious experience. How do I know that my experience of red is the same as your experience of green? Could someone who only had black and white vision know what it was like to experience purple? Those are a fact of human behavior.

There is a very interesting research project in trying to study these intuitions in adult humans, in kids, across cultures, across languages, to try and find out exactly what the data are about the puzzlement and, most interestingly, to try and find the mechanisms that generate this kind of behavior. Presumably, this is a fact of human behavior. Human behavior is ultimately explainable. It seems we ought to be able to find the mechanisms that are responsible for this expressed puzzlement about consciousness. In principle, there is a project for psychology, and for neuroscience, and for AI to try and find plausible computational mechanisms that fit the human case, explain what's going on in us so that it might have some applicability to AI as well.

You can find bits and pieces of work going on right now in psychology, in neuroscience, and philosophy that bear on this. I don't think it's yet been put forward into a research program, but I've been trying to advocate for that lately because it's a tractable bit of the mind-body problem we can bite off. The thing that makes it tractable is it's ultimately a bit of behavior that

we can operationalize, that we can begin to try to explain, which is notoriously hard to do for consciousness in general.

There are people who work on so-called "artificial consciousness," trying to produce consciousness in machines, but the whole question of criteria is very difficult in this case. In the human case, for neuroscience and psychology, you start with a human who you know is conscious and look for the neural correlates of consciousness and potential mechanisms. In AI systems, however, you don't start with a system that you know is conscious. It's very difficult to know what operational criteria you want to satisfy in order to count the system as conscious.

So, here's a potential operational criterion in something like expressed puzzlement about consciousness of the kind that we do. Once you've got an AI system that says, "I know on principle I'm just a bunch of silicon circuits, but from the first-person perspective, I feel like so much more," then maybe we might be onto something in understanding the mechanisms of consciousness. Of course, if that just happens through somebody programming a machine to imitate superficial human behavior, then that's not going to be so exciting. If, on the other hand, we get there via trying to figure out the mechanisms which are doing the job in the human case and getting an AI system to implement those mechanisms, then we find via some relatively natural process, that it A) finds consciousness in itself and B) is puzzled by this fact. That would at least be very interesting.

Will every possible intelligent system somehow experience itself or model itself as having a mind? Is the language of mind going to be inevitable in an AI system that has some kind of model of itself? If you've just got an AI system that's modeling the world and not bringing itself into the equation, then it may need the language of mind to talk about other people if it wants to model them and model itself from the third-person perspective. If we're working towards artificial general intelligence, it's natural to have AIs with models of themselves, particularly with introspective self-models, where they can know what's going on in some sense from the first-person perspective.

Say you do something that negatively affects an AI, something that in an ordinary human would correspond to damage and pain. Your AI is going to

say, "Please don’t do that. That’s very bad." Introspectively, it’s a model that recognizes someone has caused one of those states it calls pain. Is it going to be an inevitable consequence of introspective self-models in AI that they start to model themselves as having something like consciousness? My own suspicion is that there's something about the mechanisms of self-modeling and introspection that are going to naturally lead to these intuitions, where an AI will model itself as being conscious. The next step is whether an AI of this kind is going to naturally experience consciousness as somehow puzzling, as something that potentially is hard to square with basic underlying mechanisms and hard to explain.

I’m not going to say that it’s inevitable that an AI system will experience itself this way and make these reports. After all, there are plenty of humans who don’t make these reports. But in humans there are at least some underlying mechanisms that tend to push people in the direction of finding themselves to have these weird and interesting mental phenomena, and I think it’s going to be very natural for AIs to do that as well. There is a research project here for AI researchers, too, which is to generate systems with certain models of what’s going on within themselves and to see whether this might somehow lead to expressions of belief in things like consciousness and to express puzzlement about this.

So far, the only research I know in this direction is a little project that was done last year by a couple of researchers, Luke Muehlhauser and Buck Shlegeris. They tried to build a little theorem prover, a little software agent that had a few basic axioms for modeling its perception of color and its own processes. It would give you reports like, "That’s red of such-and-such a shade," and it would know it could sometimes go wrong. It could say, "I’m representing red of such-and-such a shade," and from a certain number of basic axioms they managed to get it to generate a certain amount of puzzlement, such as, "how could my experience of this redness be the same as this underlying circuit?"

I’m not going to say this very simple software agent is replicating anything like the mechanisms of human consciousness and our introspective access to it. Nonetheless, there is a research project here that I’m encouraging my friends in AI to look at with the help of our friends from psychology, neuroscience, and philosophy.

At the end of the day, of course, what does all this mean? Let's say we do find the mechanisms that generate our reports of being conscious and our puzzlement about consciousness, will that somehow dissolve the whole problem? Someone like Dan Dennett would certainly want to take that line. It's all a big illusion in explaining these mechanisms. You'll thereby have explained the illusion and explained away the problem of consciousness.

That's one line you can take, but you don't have to take that line for this meta-problem to be interesting. You could be purely a realist about consciousness in the philosopher sense, holding that consciousness is real. These reports are a fact of human behavior, and there are going to be mechanisms that generate them. If you're a realist about consciousness, as I am, then the hope is going to be that the mechanisms that generate these reports of consciousness and this puzzlement about it are also going to be very deeply tied to the mechanisms of consciousness itself.

I see this as a challenge for theories of consciousness, and there are a million of them out there. Maybe it's information integration, maybe it's a global workspace, maybe it's quantum this and that. For your theory of consciousness to be plausible, there's got to be some plausible story you can tell about why that proposed mechanism of consciousness itself would also potentially play a role in generating our reports of consciousness, because otherwise it would just be bizarre that the reports would be independent of the phenomenon itself.

It's not clear to me that many current theories meet this standard. Looking at, say, information integration theories, it's not clear to me why those theories where more and more information is integrated is likely to dispose a system to make these reports, and it looks like the reports can disassociate from the information integration in various, interesting ways. So, I see this at least as a challenge for theories of consciousness, as well as a challenge for AI research and for philosophy.

* * * *

RODNEY BROOKS: This seems not so much meta as hyper. It's a list procedure. Hyper is the next key after meta. I haven't read enough of your

writings to know whether you believe that mammals have some level of consciousness.

CHALMERS: I do.

BROOKS: I'm guessing you wouldn't expect a dog to be able to report on its own consciousness. So, isn't this a high bar for consciousness, if you're wanting it to report on itself?

CHALMERS: I don't think anyone should propose reports as a necessary condition for consciousness, clearly. Most of the time we're conscious and we're not reporting. Kids are presumably conscious well before they can report.

BROOKS: What age do kids start reporting on consciousness? Do you have any idea?

CHALMERS: It depends where you count. Are you talking about consciousness in general, the abstract category? This comes relatively late. What age do kids start talking about pain?

ALISON GOPNIK: If you're talking about things like differences between mental states and physical states, by the time kids are three they're saying things like, "If I'm just imagining a hotdog, nobody else can see it and I can turn it into a hamburger. But if it's a real hotdog then everybody else can see it and I can't just turn it into something else by thinking it." There's a bunch of work about kids understanding the difference between the mental and the physical. They think that mental things are not things that everybody can see, and that you can alter them in particular kinds of ways, whereas physical things can't, and that's about age three or four.

There is a whole line of research that John Flavell did, where you ask kids things like, "Ellie is looking at the wall in the corner, are things happening inside of her mind?" It's not until about eight or nine, until late from a developmental perspective, that they say something's going on in her mind when she's sitting there and not acting.

You can show that even if you give the introspective example; for example, if you ring a bell regularly—every minute the bell rings—and then it doesn't, and you say to the kid, "What were you thinking about just now?" The kids say, "Nothing." You ask them if they were thinking about the bell and they just say no. There's a lovely passage where a kid says that the way your mind works is there are little moments when something happens in your mind, you think, and then nothing happens in there. Their meta view is that it's consciousness if you're perceiving, or acting, or imagining to a prompt. But if you don't, if it's not connected, then nothing is happening. So, they have a theory of consciousness, but it looks like it's different.

CHALMERS: It's important to separate intuitions about mind and consciousness, in general, from intuitions about specific phenomena like feeling pain, seeing colors, or thinking. It's probably the case that intuitions about the specific phenomena in kids will kick in a lot sooner than the expressions about the category of mind or consciousness, in general.

NEIL GERSHENFELD: What do you think about the mirror tests on elephants and dolphins for sense of self?

CHALMERS: Those are potential tests for self-consciousness, which, again, is a high bar. There are plenty of animals that don't pass them. So, are they not self-conscious? No. They're probably just not very good with mirrors.

GERSHENFELD: But do you think that's a falsifiable test of sense of self?

CHALMERS: That's pretty good evidence that the animals who pass it have certain kinds of distinctive self-representations, yes. I don't think failing it is any sign that you don't. I would also distinguish self-consciousness, which is a very complicated phenomenon that humans and a certain number of mammals may have, from ordinary conscious experience of the world, which we get in the experience of perception, of pain, of ordinary thinking. Self-consciousness is just one component of consciousness.

CAROLINE JONES: I want to tie it together to Rod's question, because the question of reporting and the question of the self are distinct. One of my running thoughts was about this question of the human who has programmed the computer to report. When my car says low battery, is it

aware that it's feeling low battery? No. I've just programmed it to tell me that it needs care. I want to just propose to you the concept of self-care. When the human feels pain, it doesn't need to tell anyone else what happened.

I wonder if that could be a contribution to the engineering of consciousness in the AI that it forgets about the human that it's been told to report to and instead says, "My battery is feeling kind of low. What can I do about it?" I wonder if that model of interiority—where you self-talk, you self-report, you self-engineer, you perform some sort of self-action—would be the human model that matters.

CHALMERS: Some kind of connection to your own drives and your own self-concern?

JONES: Right. In other words, what I gathered from the book is that there are forms of AI that are beginning to self-generate self-reports and self-repairs.

GOPNIK: Even simple systems do that. Essentially, anything that's even faintly complex is going to be regulating its own operations.

JONES: I guess I'm recommending to the philosophers that they question their own paradigm of engineering this reporting mechanism.

GOPNIK: But it's not the reporting mechanism. The AI is doing exactly what you describe: "Here's an error. I've got some evidence that I'm making an error, so I'm going to modify what I do based on that."

CHALMERS: We're not yet at that level of mind and mental vocabulary. For mental vocabulary to kick in, it's probably going to have to be embedded in the systems of believing, desiring, valuing, pursuing goals, perceiving, which goes on in humans.

GOPNIK: Here's a proposal, David, that's relevant to kids not wanting to go to sleep. One of the things that's very characteristic of kids, including babies from an early age, is that at a point when they clearly have an incredibly strong drive to go to sleep, they don't want to go to sleep. If you talk to

kids, even little kids, it’s very hard not to conclude that the reason they don’t want to go to sleep is because they don’t want to lose consciousness. It’s sort of like, "I’ve only been able to do this for two years, I really don’t want to stop." I don't know whether other creatures share that.

CHALMERS: That's an intuition about the idea of consciousness, that it does something special that gives your life value.

GOPNIK: Nick Humphrey has an interesting proposal along these lines that it’s connected to things like not wanting to die, that that's the reason for the meta-intuition.

CHALMERS: So, he thinks that actually generates the problem of consciousness, because we don't want to die.

FRANK WILCZEK: We know we go to sleep, but we’re not so sure we’re going to wake up.

IAN MCEWAN: I have a constant discussion going on between my Adam and my narrator. Adam has particularly interesting eyes—blue with little vertical black rods—and every time my narrator is talking to Adam, he’s looking into these eyes, wondering whether Adam can see in the sense that we see. In other words, are his eyes functioning like cameras? Does he see like a camera sees? And that’s just a metaphor. Does he hear like a microphone hears? He poses himself the question, who is doing the seeing? But as soon as he asks himself that question, he has to pose the question of his own methods. Who is doing my seeing? There isn’t a homunculus sitting up there seeing, because a homunculus would have to have someone inside himself to see what the homunculus sees. Obviously, this was dealt with at length in the 17th century and disposed of.

Finally, they agree that what they share at the root of their consciousness is matter. The narrator has neurons, Adam has a whole set of other replicates for them, but upstream of both is the nature of matter, and they just have to leave it there. It can go no further than this.

CHALMERS: I do think, at least sociologically, when it comes to the creation of AI, this question is going to become a practical one once there are AIs in

our midst. People are going to have arguments about whether they're actually conscious. The mere fact that they can see and that they can talk about what they're seeing, all of that will help a little bit, but that won't be enough to convince many people that these are conscious beings. Once you get AIs that seem to care about their consciousness to the point where they're saying things like, "Please don't turn me off, even for a little while," or where they start experiencing puzzlement about their consciousness, saying, "I know in principle I'm just a mechanism, but I experience myself like this," these carry sociologically significantly weight in convincing people that these are conscious beings with morals.

GOPNIK: Nick has some examples with primates doing things like taking a rock and holding it underneath water and looking and feeling the water on the rock as something that evidently primates do, where it's very hard to see what functional significance it has other than valuing the experience of feeling their hand in the water and having the rock. If those things developed spontaneously, that might be an interesting way of thinking about it.

CHALMERS: The signs of enjoying your experience, the feeling that this is what makes my life worth living.

SETH LLOYD: One thing that comes across from both your talk and the discussion afterwards is there are many different kinds of consciousness. Might it be useful to simply declare that there is not one thing we call consciousness?

I had a conversation about consciousness with an anesthesiologist and she pointed out that if you're an anesthesiologist, consciousness is definitely not one thing because you have to have four different drugs to deal with the different aspects of consciousness that you wish to disable. You have one to just knock people out. It's known that people can still experience things and still experience pain, so then you have another to block the sensation of pain. People could still have memories while they're knocked out and not feeling pain, so you have to give them another one to knock out the memories that you have. Sometimes they give you an extra special one to make you feel good when you wake up. So, each of these drugs are quite

different from each other, with different functions, and they're disabling different aspects of the things that we call "consciousness."

CHALMERS: The philosophers' age-old move here is to make a distinction. I didn't want to get too much into the jargon here, but in the philosophy and the science of consciousness, there is fairly standard language by now. You separate the various forms of consciousness. For example, there's phenomenal consciousness—the raw experience; access consciousness, which is a matter of accessing things and using them to control behavior and lay down memories; reflective consciousness—reflecting on your own mental states; and, indeed, self-consciousness—consciousness of yourself. Those distinctions do need to be made. The kind of consciousness I tend to focus on the most is phenomenal consciousness—the role of experience. Even then, of course you can start breaking it down into components, so there's sensory consciousness, there's cognitive consciousness, there's affective consciousness. Don't get me started on the distinctions. I agree, there are plenty of them to make.

In fact, the anesthesia question is very interesting because it sure looks like what's doing the heavy lifting in a lot of cases with anesthesia is scary stuff like the amnestics—the things that block your memories. They've been doing a whole lot of the heavy lifting, and maybe some analgesics that block the feeling of pain, and certainly the paralytic that blocks your movements. But do any of those things actually prevent you from being conscious?

JONES: The most significant one for me is the one they give you so you don't care. There's a whole body of surgery that you're completely alert, but what they've given you is so you don't care. It's a very strange feeling. It's all going on, there's even a little pain, but you just don't care. I don't know where the philosophers put that. Does it fall into the affective subset?

CHALMERS: I'd say it's affective because it's a value, experiencing values and goals. But also agentive consciousness, which is the feeling of action. You're no longer acting.

BROCKMAN: At the Om Conference in 1973, which was perhaps the first post-Macy Cybernetic Conference—Bateson and von Foerster were the

organizers—John Lilly addressed this. He said, "The way you deal with inhibiting consciousness is very easy: baseball bat."

PETER GALISON: It's interesting because when you break it down, we can see that some of these aren't a worry we would have about getting machines to do, like not laying down memories, that doesn't sound like a hard thing to model with the machine, or paralysis, being unable to effectuate some motor or prosthesis or something, that doesn't seem like a hard thing to put into a machine.

The advantage of the kind of distinctions that you were just making is that it then isolates the part that seems weird and troubling to us. When we say, "Machines don't have consciousness," we certainly don't mean machines can't lay down memories or machines get paralyzed so they can't affect their motor actuators. It's something like the self-aware component.

CHALMERS: I want to say it's the phenomenal consciousness component, the raw, subjective experience, which may involve self-awareness, but I'm not sure it has to. If it turns out a machine is experiencing pain and having a visual experience with the world, of the kind we do, that would be remarkable. That's part of what we care about in machine consciousness. Certainly, the one that seems the most puzzling to me is not actually self-consciousness, per se, it's just straight up subjective experience.

GALISON: So, you would think of a squirrel having pain?

CHALMERS: Yes. A squirrel almost certainly has some kind of subjective experience. The question is at what point are AI systems going to have that?

BROOKS: A few minutes ago, you were talking about this becoming a real issue when we have artificial intelligent systems around, but it becomes an issue much earlier than that because people's attribution leads them in strange ways. We saw this in my lab in the '90s with people interacting with the COG robot with Kismet robot.

CHALMERS: By the way, there are a lot of psychological results that show the number one thing that convinces us that a system is conscious is

whether it has eyes. So, you go through a whole bunch of different systems and if they have eyes, they're conscious.

WILCZEK: Are you a vegetarian?

CHALMERS: I'm not. I used to think I shouldn't eat anything that was conscious, but my views are such that consciousness is very widespread in the animal kingdom and possibly outside, so I'd likely go very hungry. There's a lot of research now on the impressive things that plants can do.

MCEWAN: We would find it very hard not to attribute a being with consciousness if it appears to have a theory of mind and appears to understand us.

CHALMERS: Maybe there's an AI that mimics certain superficial behaviors. I'm thinking of a little cartoon AI who's studying up for the Turing test, and it reads the book called *Talk Like a Human.* Maybe superficially he could get one or two sentences in to convince us he's conscious, but in order to mirror all of our sensitivities and our expressions of the varieties of consciousness, the project is not just to mirror superficial expressions, but to mirror the underlying mechanisms. Once I have an AI based on the mechanisms in humans and they give rise to the full range of expression, I'm not sure how much more I could demand.

BROOKS: My early experience in the '80s when I was building insect-like robots was all about the speed. So, if the robot just banged into the wall and backed up and did it again slowly, people would ask what was wrong with it. But if it did it fast, they would say it looks frustrated.

JONES: In 1943, Fritz Heider and Marianne Simmel put this into their short animated film.

The cyborg interface is something that's got to come into the futurology here, because if I'm plugging in an infrared sensor and then I share it with my computer and we have a certain phenomenal platform between us, at what point is my consciousness circuiting? At what point do I deposit some of my reflective capacities into the device, having shared certain machinate possibilities and so on and so forth. This goes to Frank's beautiful concept of

the evolving ecology. We are persisting in thinking of this "other" as a heap of metal that is going to somehow eventually arrive. But what if we are tutoring it, what if we are participating and trading its perception to our perception, then parking it when we go to sleep? That's a possibility that philosophers could help us imagine because it's already happening.

CHALMERS: We've already offloaded a lot of our cognition into our devices—memories, planning, and navigation.

JONES: There's an artist who has planted a thing in his brainstem so that he can hear colors because he's colorblind. What part of his consciousness of colors is in the chip, in his cochlear enhancement device? These questions are already evolving in our partnerships with machines, so we might as well think about whether we're going to take a pedagogical position in relationship to that.

CHALMERS: Especially once there are serious brain computer interfaces. This is going to be the point where consciousness starts to extend into our devices.

JONES: The question is whether the wild child of Aveyron had consciousness, right? There was no human to say, "Are you in pain? Oh, are you hungry? Is that your internal state?" That's a pedagogical environment that nurtures and teaches and evolves consciousness. So, I think we could do that with machines.

GEORGE DYSON
AI That Evolves in the Wild

I'm interested not in domesticated AI—the stuff that people are trying to sell. I'm interested in wild AI—AI that evolves in the wild. I'm a naturalist, so that's the interesting thing to me. Thirty-four years ago there was a meeting just like this in which Stanislaw Ulam said to everybody in the room—they're all mathematicians—"What makes you so sure that mathematical logic corresponds to the way we think?" It's a higher-level symptom. It's not how the brain works. All those guys knew fully well that the brain was not fundamentally logical.

We're in a transition similar to the first Macy Conferences. The Teleological Society, which became the Cybernetics Group, started in 1943 at a time of transition, when the world was full of analog electronics at the end of World War II. We had built all these vacuum tubes and suddenly there was free time to do something with them, so we decided to make digital computers. And we had the digital revolution. We're now at exactly the same tipping point in history where we have all this digital equipment, all these machines. Most of the time they're doing nothing except waiting for the next single instruction. The funny thing is, now it's happening without people intentionally. There we had a very deliberate group of people who said, "Let's build digital machines." Now, I believe we are building analog computers in a very big way, but nobody's organizing it; it's just happening.

GEORGE DYSON is a historian of science and technology and author of *Darwin Among the Machines* and *Turing's Cathedral.*

* * * *

GEORGE DYSON: I'm not a scientist. I've never done science. I dropped out of high school. But I tell stories. Ian tells stories that can take us into the future wherever he wants to go, and I go into the past and find the stories that people forgot.

Alison Gopnik said how nobody reads past the one sentence in Turing's 1950 paper. They never read past his 1936 paper to his 1939 "Systems of Logic

Based on Ordinals," which is much more interesting. It's about non-deterministic computers, not the universal Turing machine but the second machine he wrote his thesis on in Princeton, which was the oracle machine—a non-deterministic machine. Already he realized by then that the deterministic machines were not that interesting. It was the non-deterministic machines that were interesting. Similarly, we talk about the von Neumann architecture, but von Neumann only has one patent, and that patent is for non-von Neumann architecture. It's for a neuromorphic computer that can do anything, and he explains that, because to get a patent you have to show what it can do. And nobody reads that patent.

The measure of a good story is that it gets better as it's repeated by other people, such as Danny's story about the Songs of Eden and how you can look at the development of language and consciousness from the point of the view of the songs themselves, these strings of language. We're obsessed with these other minds that are going into technology. There's a whole other track where you could have a mind and intelligence that has no technology at all. Freeman always pointed out that the search for extraterrestrial intelligence is wrong, that really what we are looking for is extraterrestrial technology because we can see it. Intelligence and technology are different things. There's a parallel to the songs that went to the apes becoming us, and the songs that went into the oceans and became whales, which have highly developed songs and are raised by their maternal 100-year-old grandmothers. Whales have no technology, but obviously they have very advanced brains, five, six, eight times the size of ours.

I'm interested not in domesticated AI—the stuff that people are trying to sell. I'm interested in wild AI—AI that evolves in the wild. I'm a naturalist, so that's the interesting thing to me. Thirty-four years ago there was a meeting just like this in which Stanislaw Ulam said to everybody in the room—they're all mathematicians—"What makes you so sure that mathematical logic corresponds to the way we think?" It's a higher-level symptom. It's not how the brain works. All those guys knew fully well that the brain was not fundamentally logical.

We're in a transition similar to the first Macy Conferences. The Teleological Society, which became the Cybernetics Group, started in 1943 at a time of transition, when the world was full of analog electronics at the end of World

War II. We had built all these vacuum tubes and suddenly there was free time to do something with them, so we decided to make digital computers. And we had the digital revolution. We're now at exactly the same tipping point in history where we have all this digital equipment, all these machines. Most of the time they're doing nothing except waiting for the next single instruction. The funny thing is, now it's happening without people intentionally. There we had a very deliberate group of people who said, "Let's build digital machines." Now, I believe we are building analog computers in a very big way, but nobody's organizing it; it's just happening.

If you look at the most interesting computation being done on the Internet, most of it now is analog computing, analog in the sense of computing with continuous functions rather than discrete strings of code. The meaning is not in the sequence of bits; the meaning is just relative. Von Neumann very clearly said that relative frequency was how the brain does its computing. It's pulse frequency coded, not digitally coded. There is no digital code.

In mathematics there's this deep, old problem called the continuum hypothesis. We have an infinite number of different infinities, but they divide into only two kinds: countable infinities and uncountable infinities. My analogy for that is how at the end of a conference when you look for a t-shirt, there are only extra small t-shirts and extra large. There are no medium t-shirts. The continuum hypothesis—and there is a difference between being true and being provable—has not been proved. It says you will never find a medium-sized infinity. All the infinities belong to one side or the other.

Two very interesting things are happening. What this means is that for any uncountable infinity, say, a line, there's an infinite number of points between any two points, and then if you cut a piece of that line, it still has an infinite number of points. That, I believe, is analogous to organisms. All organisms do their computing with continuous function. In nature we use discrete functions for error correction in genetics, but all control systems in nature are analog. The smallest analog system has the full power of the continuum.

On the other side, you have the constructible infinities. What's interesting there is that we're trying to prove this by doing it. We're doing our best to

create a medium-sized infinity. So, you can say, "Well, it exists. We've made it." The current digital universe is growing by 30 trillion transistors per second, and that's just on the hardware side, so we have this medium-sized infinity, but it still legally belongs to the countable infinities.

My metaphor of how I think about this is that no matter what you do in the digital world, it stays stuck on that side of the room. But there's no prohibition against machines doing continuous computing. Then they belong to the other side. We were talking about hybrid machines yesterday. That's the interesting future that the Adam that Ian McEwan imagines is only going to happen when the machines move to the other side, to the continuous side. Then they can start having the things we have. There's no reason not to do that.

I'm just going to close with not my idea but somebody else's. The von Neumann centennial was in 2003, and the Templeton Foundation was changing from trying to prove the existence of God to not mentioning God at all. They held a series of meetings in honor of von Neumann, one of which was on von Neumann game theory. One of the people, a Scottish mathematician, came in and gave absolutely beautiful proof using classical von Neumann game theory. It wasn't proof of the existence of God, but it was proof that if there was a God, no matter what value function you choose, the payoff is higher if God does not reveal herself.

The message to take home is that faith is better than proof. You don't want proof. We're in exactly the same situation with AI. We have these meetings year after year with the same discussions, and people are waiting for proof. To me the Turing test is wrong. Actually, it's the opposite. The test of an intelligent machine is whether it's intelligent enough *not* to reveal its intelligence. It's true for AI as a whole that we're going to keep coming back, and we need to have faith in AI. I have faith in it. I believe it exists, but we don't want proof. It's a game of faith.

* * * *

W. DANIEL HILLIS: George, I wonder if you're making too much of this distinction between continuous and discrete.

G. DYSON: Oh, I'm definitely making too much of it.

HILLIS: To me there's an engineering problem in systems, which is caused by noise. Analog systems generally deal with that problem by filtering. So, they do it by only accepting restricted time frequency range of signals. In some sense they disallow information being encoded in a certain part of the frequency space. Sometimes that's just inherent in how they're built. Sometimes it's done by vector explicit filtering.

Another way of dealing with noise is disallowing certain amplitudes, which is basically how digital systems do it. That has some advantages and disadvantages. Either of them can be made to represent things to arbitrary precision, and in practice you can represent things to higher precision using the digital methods, although at great cost in power.

So, it seems to me like this is just an engineering trick. There are many other things that are halfway in between, like using the discrete eigenvectors of a continuous function or something like that. It seems to me like there's nothing qualitatively different. It's an interesting engineering discussion to say, "Hey, I might do better using analog to solve these problems," but I'm not sure that in terms of its ability to do an artificial intelligence that there's anything there.

NEIL GERSHENFELD: On the analog side, you can price that exactly with fluctuation dissipation. There's an exact pricing for the thermodynamic cost of having tolerance in an analog signal. The very first digital logic had floating point processors, and they had digital signal processors, and they had digital signals. They had processors to do special processing on continuous costings on the digital side. On the analog side, there's a very precise tradeoff between what tolerance costs you, and in fact most of the power in your phone's radio is in the receiver, not in the transmitter against this fluctuation.

G. DYSON: Analog machines, like nervous systems, don't have the programming. There's not an algorithm, which is where we're wrong. We're so obsessed that there has to be an algorithm.

RODNEY BROOKS: Instead of worrying about whether it's analog or digital, it's the organization, because you get into a different computational complexity class by the way stuff is organized.

HILLIS: That's the second sense of analog. There are two completely different senses of analog, which have nothing to do with each other.

BROOKS: He was talking about your second sense, I thought.

HILLIS: I thought he was explicit that he was talking about continuous versus discrete.

G. DYSON: Yes. I didn't get to the other side, which is that nature builds very intelligent systems without any digital programming in the sense that we take it for granted.

HILLIS: Then there's a second sense of analog, which is in some sense whether the computation bears an analogous structure to the thing that it's computing on. For instance, a map is an analog of the physical. You can have continuous and discrete circuits that are analog in that sense, that work by an analogy in the world.

Having the algorithms stored separately versus inherently built into the structure is yet another issue. We tend to talk about all those together, and they get mixed up in this digital/analog distinction. I'm not sure what the interesting distinction is.

GERSHENFELD: To Rod's point, these are ridiculous extremes. If analog is the needle on the DVM and digital is ones and zeroes, neither really bears on what's interesting. Both in biology and in computers, salvation is in the details and the architecture, which applies a really interesting space that's not captured by either of those limits.

SETH LLOYD: Historically, this whole question was the subject of Shannon's great book, *The Mathematical Theory of Communication*, where he showed exactly how if you have analog systems, continuous systems that have noise and that are power and bandwidth-limited, then they are effectively digital, and you can map the number of bits that can be encoded in it. This is the

book where he coined the word "bit," which he stole from John Tukey. In some sense this is a question that was resolved gloriously in 1946.

BROOKS: A few years ago, Carver Mead told me that the defining moment of his life was when Gordon Moore handed him a bag with thirty transistors in it.

G. DYSON: He wrote the book on analog VLSI!

HILLIS: Freeman made an engineering observation that we've gone overboard with this digital thing, and it's very costly. It's probably not the right technology to get to the next level of performance. These things would be better done using analog. I agree with that point that we've over-pushed the digital thing in our engineering. But that's an engineering technology point; it's nothing fundamental about computation. I thought when you started making this analogy with the continuum hypothesis that you were saying there's some fundamental difference between these computations. I don't believe that one.

G. DYSON: The analogy was that when you take the continuous infinity and cut it in half, you still have the full infinity. The two kinds of computing follow the same path.

HILLIS: Here's why that's not true: If you cut the analog signal in half, you've now got twice as much noise per signal.

GERSHENFELD: Fluctuation dissipation means if you multiply how much a signal fluctuates by how much power you're consuming based on the system you're in, that's a constant, and so reducing the fluctuation proportionately increases the power consumption. It costs you to limit fluctuation in analog systems. They're not continuous. It's very expensive to bound the distribution. The people making all the devices around us live in that. It's this naïve version of this beautiful, clean dot on the line.

TOM GRIFFITHS: You can see a nice example of this in human languages. They way that human languages are structured there is through a continuous signal that is coming out of our mouths. The way that we perceive that is by breaking it up into these discrete paths, phonemes and

so on, and then building those into words and then being able to exploit the common atorics of the resulting discrete signals.

ROBERT AXELROD: Intonation is analog, right?

GRIFFITHS: Yes, but that's layered on top of an underlying digital thing. There's a nice experiment that was done by Simon Kirby and his colleagues where they had people playing slide whistles and then asked people to reproduce the slide whistle sounds, and then they looked at what happened as those slide whistles were transmitted. They very quickly evolved into discrete digital signals of repeating particular elements and so on. The argument is that that's essentially what happens in language evolution, too, where you get this discreteness emerging as a way of dealing with this noisy continuous signal.

CAROLINE JONES: I'd like to reorganize the discussion to his last point, which was about faith, and ask if you are contesting Wiener's metaphor that John kept throwing at us about kissing the hand that holds the whip. Just what are you articulating here, that we should have faith in the self-organizing benignity of AI?

G. DYSON: No. Lack of proof is not proof of lack of existence. Just because people are saying, "Oh, we don't think there's real AI because we don't have proof of it," my faith is different. I'm quite willing to believe in it without needing proof.

JONES: So, you're advocating faith without worship.

G. DYSON: Yes. I'm just as suspicious as Norbert Wiener. In fact, I'm more suspicious than Norbert Wiener. What he was talking about was if you hand this over to the corporations, you're in trouble.

Wiener was very preoccupied with control and control systems. Now, we talk much more about intelligence. We talk less about control. Control is just as important, and there again is my faith that these large analog control systems are—that works both ways because they're not programmed. There is no program for an analog control system in the sense that you can change a bit here and get a different outcome. That's the way the world works, and

that's why we're fooling ourselves by thinking that there is somewhere a program that has control.

ALISON GOPNIK: This gets back to just how surprising it is that taking the phenomenology of verbally thinking through or calculating a process, that that very high level linguistic phenomenology, which essentially is what Turing was doing, and taking the structure of that turned out to be as productive as it was for creating—whether you call it intelligence or not—incredibly complex functions. That's a remarkable fact.

I don't think a priori if you looked at human beings and said, "Look, almost everything that's going on under the hood doesn't have these characteristics of being digitized or being sequential," and it turns out that treating that little tiny bit on top that's about how we talk to one another or how we talk to ourselves as the relevant structure turns out to create these systems that can do all things like see or process vision or create images. That's just a remarkable non-obvious scientific fact.

PETER GALISON: Early Wiener, in the war and just after the war, his interest in control, which is crucial, was attached to a notion of purposefulness. But purpose was not purely computational as such. He thought that was the leading edge of a series of analog procedures that would substitute for various mental states, a kind of post behavioral behaviorism, a behavior-accessible form that would get at a mental state. Old-style behaviorism would refuse any attribution of mental states that are useful for them, but Wiener had built on things that were going on in psychology in the late '30s. He then had this way of trying to make circuits that would do something like purposefulness and to say, "This and no other is what purpose is."

GOPNIK: There's an interesting connection there as well that the context in which you get human beings generating, and language is an interesting example, but there's at least an argument that language is parasitic on things like long-term planning. So, what's the context in which you get this phenomenon of having a series of calculations or having a series of discrete things that you're doing?

The context is things like tool use, where you have to restrict a set of actions that you're going to perform in the service of having a goal. As opposed to things like vision that don't seem to have that structure or that goal-directed teleological character. If you want to go out and see things, it's not like what you're doing is performing a whole set of operations in order be able to see something in the way that when you're saying to yourself, "Okay, what am I going to do tomorrow? I'm going to go here and I'm going to go there," has that structure. So, there might be a relationship between the idea of control and the idea of teleology and computation at least from the perspective of what human cognition is like.

PETER GALISON
Epistemic Virtues

I'm interested in the question of epistemic virtues, their diversity, and the epistemic fears that they're designed to address. By epistemic I mean how we gain and secure knowledge. What I'd like to do here is talk about what we might be afraid of, where our knowledge might go astray, and what aspects of our fears about how what might misfire can be addressed by particular strategies, and then to see how that's changed quite radically over time.

~~

James Clerk Maxwell, just by way of background, had done these very mechanical representations of electromagnetism—gears and ball bearings, and strings and rubber bands. He loved doing that. He's also the author of the most abstract treatise on electricity and magnetism, which used the least action principle and doesn't go by the pictorial, sensorial path at all. In this very short essay, he wrote, "Some people gain their understanding of the world by symbols and mathematics. Others gain their understanding by pure geometry and space. There are some others that find an acceleration in the muscular effort that is brought to them in understanding, in feeling the force of objects moving through the world. What they want are words of power that stir their souls like the memory of childhood. For the sake of persons of these different types, whether they want the paleness and tenuity of mathematical symbolism, or they want the robust aspects of this muscular engagement, we should present all of these ways. It's the combination of them that give us our best access to truth."

PETER GALISON is a science historian; Joseph Pellegrino University Professor and co-founder of the Black Hole Initiative at Harvard University; and author of *Einstein's Clocks and Poincaré's Maps: Empires of Time.*

* * * *

PETER GALISON: I'm interested in the question of epistemic virtues, their diversity, and the epistemic fears that they're designed to address. By

epistemic I mean how we gain and secure knowledge. What I'd like to do here is talk about what we might be afraid of, where our knowledge might go astray, and what aspects of our fears about how what might misfire can be addressed by particular strategies, and then to see how that's changed quite radically over time.

The place where Lorraine Daston and I focused in the study of objectivity, for example, was in these atlases, these compendia of scientific images that gave you the basic working objects of different domains—atlases of clouds, atlases of skulls, atlases of plants, atlases in the later period of elementary particles. These are volumes, literary objects, and eventually digital objects that were used to help classify and organize the ground objects of different scientific domains.

In the periods you might schematize by being 1730-1830—and these dates are arbitrary and overly precise—there was a desire above all to find the objects that were in back of the objects that we happen to see. In other words, not this clover outside the boardroom that's been half moth-eaten and half sunburned, but the plant form that exists behind that. That's what Goethe meant when he talked about the "urpflanze." The advantage of that seemed obvious. The fear was that you would spend your time looking at particular defective clovers here or there and not understand that they were unified under a particular form that was the reality behind the curtain of mere appearances.

When William Cheselden in 1733 hung a skeleton and looked at it through a camera obscura, he wasn't looking to draw that particular skeleton; he was trying to use that and then correct the errors—the fact that it was too fat, or too thin, or had a cracked rib. When Albinus said, "I draw what I draw and then I fix the imperfections," it was because it seemed obvious that the images you would want of a skeleton—or a flower, or an insect, or whatever it was—was not the skull that belonged to me or you, but the skull that belonged *behind* all the particular skulls that we might see.

There was a fear of the multiplied variegated skulls, or clovers, or clouds that we might see, and the antidote was to draw something abstracted from that that was supposed to lie behind any particulars. Goethe would say, "I never draw any particular thing." There was a particular kind of person who

was appropriate to doing this, and that was the genius. In the 18th century it was recognized that it was fine for an Albinus, or a Goethe, or a Cheselden to make that kind of argument.

In the 19th century that begins to proliferate. When everybody starts writing down, or drawing, or painting the objects that they thought should be there and they start to clash, there's a new problem brought about by the conflict of the myriad depictions of the heart, or the skull, or the plant world, or the natural world, or crystals, or other things. The epistemic fear was of this contradictory multiplication of representations, each of which purported to be the urpflanze or the equivalent in other domains. The response to that was to seek out mechanical transfer of the world to the page. And by mechanical, that doesn't mean just the levers and pencils, it could mean any kind of thing, including chemical-based photography. In the 19th century, mechanical meant all of those procedural developments.

This was labeled objectivity for the first time in a sense that's continuous with the modern sense. When Descartes uses a term like "objective," he means more or less the opposite from what we do, but that's another story. Starting around 1830, coming from a mix of literary and scientific sources, people start to talk about this as the mapping of the clover to the page—whether it's tracing, or rubbing, or photographic representation—to minimize our intervention.

If Goethe, Cheselden, and Albinus were maximizing our intervention because they were the sort of people who could part the curtain of experience, the 19th century wanted to minimize that because people didn't trust the multiplied number of scientists in the world. They wanted to know what was actually there—the skull of this person in Case 23 in the Museum of Natural History in Berlin. So, that became a different response to a different fear that had swung the other way.

Then a new kind of problem arose in which there were lots of different skulls, each correctly or isomorphically represented, at least that was the ambition. People began to question how we know whether a skull has a tumor or whether it's just a normal variation. So they started to have atlases of normal variations. You can see how this leads to a regressive problem that could go on forever, because the space of possible variations of skulls

just within this room is pretty large. Now think about extending that over all of humanity and all of time. It became very hard to work.

The way the doctors used these atlases was to identify what's normal and what's within the range of normal, so if they saw something that didn't look like that, it was pathological. What got me interested in this in the first place were these atlases of cloud chamber and bubble chamber images in particle physics, where it was used in a very interesting way. This is a literary form that the physicists borrowed from the doctors, even though physicists don't like borrowing from doctors. They said, "If you see an image that departs from the range of the normal, what you have is a discovery, not a pathology." So, the bubble chamber scanners at Berkeley, or CERN, wherever it was, would study these compendia and then send an alert through the chain of command. Once it got up to a Luis Alvarez or somebody else, they could say they discovered something.

What then began to happen was people started to see the importance of using judgment. This pure mechanical objectivity was proliferating like crazy with all these variations. People needed to know the difference between a misfiring of the apparatus or the environment and what the real effect was. The people making magnetograms of the sun said, "We could print mechanically and objectively what we get out of our machines. You wouldn't be able to tell what's an artifact of our machines, but we know." The implication was not because they're geniuses, but because they were trained. That kind of trained judgment became a new objectivity.

People began to worry about how they would train people to recognize artifacts and to do it in a way that follows a course or a procedure. For instance, there is a famous atlas of electroencephalograms, and people said, "Do our course for several weeks, and we can train you to distinguish grand mal and petit mal seizures and various other things, not because you're a genius but because we can train you."

That became a mantra in the 20th century, that you have all these atlases that explicitly extoled the human capacity to learn judgment. They could train anyone to look at electroencephalograms to make these kinds of distinctions in a way that was repeatable and therefore objective, but not

mechanical. They didn't know how to do that. This is in the '40s, and '50s and '60s. They didn't know how to make it purely algorithmic.

The same was true in stellar spectra and other astronomical problems. Long before you could classify stars by a procedure or an algorithm, people became very good at classifying them by looking at the spectra and making judgments. These are shifts in response to fears. Epistemic virtue is the response; it's the Rx to the Dx. The diagnosis of the problem was some fear, and these are responses of procedure, of judgment, of mechanical transfer to those difficulties.

There's a current project that I'm involved with in various ways, the Event Horizon Telescope. They're trying to make images of very distant objects, like supermassive black holes and other objects in the sky. One of the problems is that the data is extremely sparse and noisy, and you have to extract an image from it.

There are two problems, one of which is the spring of Narcissus problem. The spring of Narcissus problem is that you can't just print what you see because you don't see anything. If I gave you a bunch of points and told you to draw the best curve through it, you would rightly tell me that's not a well-posed question. You can say, draw the best straight line through it, and that would be easy—every ninth grader can do that. If you want the best circle or the best hyperbole, whatever it is, you can solve the problem. You need to assume something, then you can get information out. These images have that character. You have to make some kind of Bayesian assumption prior, and then from that you can create an image. That leads to the problem of Narcissus. The worry is that you might impose so heavily your prior assumption about what you would see that you would see it when it wasn't there, like when Narcissus looks into the spring and sees his face. If you don't impose any prior knowledge, then you have the opposite problem—the problem of the helm of darkness, which means you that don't see anything, so you can't extract anything. You have this sparse, noisy image.

The question on this collaboration is, how do you get an objective image? There are various strategies that they've taken. They're very interesting. For instance, one of them is to divide the team of about 120 or 200 people on this collaboration, but the imaging teams are divided into groups and they

work under utter secrecy from each other within the collaboration. They produce their images and compare them. Another strategy is to vary the priors, and then the question is, have you varied the priors enough to give you an objective image?

There’s another possibility, which has been suggested by the AI folks. For example, the space telescope has a huge number of galaxies, more than the astronomers could cope with, to try to classify and understand. One of their first moves was to make this into a public game. There are hundreds of thousands of people who do this thing called "Galaxy Zoo," where you’re given images, you take a training program, you take a test, and then you start classifying galaxies. Some people didn’t like that, so they suggested training computers to classify these galaxies. So, they began to train the computer to classify the galaxies using these learning neural net arguments. So then they said, "Okay, we’ve classified these things, which is great, but we can’t interrogate the program as to what it’s doing." It had that obscurity that we’ve talked about here before. This is a different kind of problem, where you've gained opacity and capacity at the same time. You can classify a lot of things, you can show it overlaps in the restricted domain where you’ve got experts and it gives the right answer, but you don’t really know what it has done.

There are lots of interesting papers where people start to talk about AI and attributing to it a kind of human capacity. They say it mislearned, or started to act pathologically; it found a little bit of striping on the snail and has covered the snail completely with stripes, made it look like a zebra snail. This attribution of purpose and humanity to this program, partly in virtue of the fact that it seems to be making human kinds of errors, becomes a big issue because you can’t ask it what it’s doing. We'd like the AI to take over some of these tasks as a way of solving the objectivity problem, but then in response we have an opacity problem. That happens in a lot of domains.

In my contribution for the book, I talked a little bit about algorithmic sentencing. This is, for instance, if a judge wants to sentence people on their objective likelihood of committing another crime. The problem is that because of the proprietary secrecy of the company that makes the algorithms, or because the algorithms are so complicated that they can’t unwind them, they don’t know or won't be told how the decision is being

made and whether it's using criteria that would violate our norms. So, if you live above 125th Street in Manhattan and you're given a higher sentence, this is just a proxy for race.

That's in the moral, political, legal domain, but in the epistemic domain of the sciences, there are analog questions that you might ask. What kinds of criteria are being emphasized in this? What is the alternative to this opacity? We know what the gain could be: It could increase our capacity, give us objectivity beyond human judgment. But it costs us in what we can interrogate. Suppose that it worked, suppose we were completely happy with it. Would that be enough in the scientific applications of AI? I don't mean what you buy on Netflix or Amazon; I certainly don't mind not knowing the algorithm by which it tells me I might like a movie if I like the movie. I just question whether even excellent prediction in the scientific domain would satisfy us.

I just want to end with a reflection that James Clerk Maxwell had back in the 19thcentury that I thought was rather beautiful. James Clerk Maxwell, just by way of background, had done these very mechanical representations of electromagnetism—gears and ball bearings, and strings and rubber bands. He loved doing that. He's also the author of the most abstract treatise on electricity and magnetism, which used the least action principle and doesn't go by the pictorial, sensorial path at all. In this very short essay, he wrote, "Some people gain their understanding of the world by symbols and mathematics. Others gain their understanding by pure geometry and space. There are some others that find an acceleration in the muscular effort that is brought to them in understanding, in feeling the force of objects moving through the world. What they want are words of power that stir their souls like the memory of childhood. For the sake of persons of these different types, whether they want the paleness and tenuity of mathematical symbolism, or they want the robust aspects of this muscular engagement, we should present all of these ways. It's the combination of them that give us our best access to truth." What he was talking about in some ways was himself; this is what he wanted.

If you go back to one of the great Old English origins of the word "understanding," under doesn't mean beneath, it actually meant "among." "Standing," was different forms of standing. It's almost like you're standing

in a grove of different trees. That sense of being among these different ways of grasping the world—some predictive, some mathematical—even something as abstract as a black hole, there are models that use swirling water like a bathtub around a bathtub drain to understand the dynamics of the ergosphere, that ability to stand among these different things might be something that we want and whether we can make use in different ways of AI, or whether AI will only be part of that understanding seems to me to be known.

* * * *

NEIL GERSHENFELD: Peter, there's one straightforward response. When you say the network is inscrutable, that's an early, simple version. There's an interesting thing happening with what are called auto encoder networks, where you force the network through a constriction and you force it to have low-dimensional representation after it's gone through this high-dimensional unpacking. There are been a lot of interesting results where you then look at these internal representations and find they're interpretable. It's a simple version just to say it's a big network and the output comes—you can ask the network to help you find a representation. There have been a number of interesting examples of what comes from those.

GALISON: Yes, I've seen them. There are a bunch of different ways of sampling in the space that can help you, but the people that do a lot of this imaging work find that they are unable to unwind those.

GERSHENFELD: What I'm saying isn't sampling, it's something different that's a little more recent. As part of training the network, you train the network through an internal constriction where you ask it to find an interpretable representation. So, it's a different architecture from just you look at the network and figure out what it's doing. You train the network to teach you a representation you can understand. There are very interesting examples of that working.

W. DANIEL HILLIS: I understand the sense in which you say the networks are inscrutable, but I'm surprised that you think people are scrutable. If you ask somebody why they decided something, they will make up a story.

There’s very good evidence that in many cases that story has nothing to do with what happened.

GALISON: That’s true. But you can sometimes get farther with it. At least that was the hope.

HILLIS: I would think there's a better hope, particularly if you want to have networks that have the property of understandability, which is the kind of thing Neil is talking about—to have AI that is truly understandable in how they made the decision. There's more hope with that than with humans.

ALISON GOPNIK: There are two orthogonal problems that are getting mixed up here. One of them has to do with how much access the system has to its own processes. The other one, which is the scientific problem, has to do with whether the system is outputting a representation of what’s actually out there in the world. If you think that all of human cognition, or at least a lot of it, is this inverse problem about a bunch of data that’s coming and you want to reconstruct what it was out there in the objective world that was creating that data, that’s the central problem of something like a visual system and it’s the central problem of science.

First of all, do you understand what the process is that is enabling you to solve that inverse problem? Secondly, do you have something that looks like a solution to the inverse problem? Do you have a representation, whether it’s accurate or not, about what’s going on in the world outside that’s leading to that pattern?

HILLIS: To answer that second question, though, you have to have some criterion of the quality of the solution. That’s a very well studied thing in classifier theory. There are many measures of the quality of the solution when you decide to basically cluster things. So, you can pick your measure, and you can measure how good it is under that measure.

GALISON: It's not that people said, "Humans make judgments. That’s fine." In fact, what led to the development of mechanical objectivity in the first place was that they didn’t like relying on those judgments, even by people like Goethe and Albinus, because they felt that it was obscure. When it began to be a real problem was when it proliferated in the 19th century.

IAN MCEWAN: In your early atlases, we see the power of Platonic thought and the extension of Neoplatonism.

GALISON: Yes, that's how I think of it. It's a sense that you can find a pure form that lies behind the myriad particularities that we encounter.

MCEWAN: This is Plato's cave, in fact.

HILLIS: There is a way of representing clusters, which is to pick the center of the cluster and then pick all of the things that are closest to it. The better algorithms like vector support methods pick a bunch of outliers, so anything farther out than this won't be considered in the cluster.

GALISON: In the history of classifying images by scientists, there are a bunch of different strategies. One was to take the most perfect extant instance—the best skull—abstract from that, make it even more perfect, maybe geometrizing it in some way or making it into a perfect harmony of measures. Another was to take an extreme example or an average. There were atlases that would take like many livers and weigh them all and find the average weight of a liver, and then that became the notion. That's more like your center choice that you were saying. Then in the biological domains, very often they the first-discovered instance becomes the type specimen, which is even stranger.

GERSHENFELD: Harvard has this amazing room of drawers in the museum. So, you pull open the drawer and it will look like little fur pelts, and the fur pelts might be for a beaver and a squirrel. But it's not *a* beaver; it's *the* beaver. It is the beaver that all beavers are defined by. That's become much more important recently because they're now sequenced, and they're used to do genotypes and phenotypes.

GALISON: You can see that there's a struggle to try to figure out how to make a representation of a class of things that are different.

HILLIS: In clustering techniques, the ones that work less well in practice or the two that you mentioned before—pick the one that's closest to the center, or make up an imaginary one that's in the center—those don't work very well.

It turns out that the ones that seem to behave the best in practice is something that was not in any of your lists. I'm not sure if this is ever done in atlases, but it's basically what I call support vector, where the support vector is the set of things that are right on the edge. You define it by the things that are barely within the category.

GOPNIK: Danny, is it hat you have an objective measure of what the things are that are being clustered, independently of the cluster?

HILLIS: Yes.

GOPNIK: That's the problem that Peter is raising. When you're doing science, you're in this situation in which you are trying to do the clustering and you're trying to figure out what the thing is that's generating the data that you want to cluster.

GALISON: When you're looking at a candidate black hole with surround and you don't know what it's going to look like, it's different than saying, "There are three kinds of galaxies. I want to classify them. This one, I know it looks like one of those, etc."

CAROLINE JONES: The problem seemed to proliferate off earth. Out of Plato's cave is fine when you can wander around and pick up the turtle or the beaver, but when you're drawing the canals of Mars, you don't have the symbols, you just have geometries that it turns out you're imposing. You're imposing on the mechanical artifact.

SETH LLOYD: I disagree. Neoplatonism is never fine.

JONES: As a historical progression, Platonism works for those people who have philosopher kings and can wander and pick up the beaver. It works less well when you're relying on a telescope and looking at a surface of a distant planet.

LLOYD: You're describing support vector machines, where it's a mathematically well-defined process to make a cluster, or you have two clusters and you try to draw the hyperplane that separates them with the maximum margin, which is a good idea and works extremely well also in

high dimensions. Then, of course, the stuff on the edge of this margin, the support vectors define it. Then you could also do k-means clustering, which is the other one. You pick a represented example and you say, "We can put these together," and you find that both of them work okay, but of course, there aren't necessarily clusters. There's no definition of what a cluster is.

These things will have examples of overlap with each other, so you'll have things that impinge on the other clusters. There is no abstract ideal cluster that's there, so you have to come up with some reasonable Bayesian prior to say, "Okay, here is how we're going to deal with this situation."

Now, we have this nice feature where we admit these artificial intelligences, which we are not going to understand in roughly the same way we don't understand humans, into our spectrum of models that we're going to trust in order to do things like look at medical images. In addition to radiologists looking at medical images, we can also run them by the deep neural network and see what they say, too.

It's rather nice that we have other artificial intelligences with whom to collaborate. We don't know what they're doing, but we also have other methods that we can compare, things like support vector machines and well-defined mathematical methods where we know what's going on, which is incidentally is what happens in Netflix. Netflix does not have a deep neural network, they have a matrix completion algorithm, where it's well defined mathematically. It's very labor intensive, but we could walk through it and say, "Here's what's happening inside your computer exactly and here's why it works."

HILLIS: In many examples, there is an outside way of measuring the volley, which is if you're going to do something with the decision, it's the utility of the success of doing whatever. There are many systems in which you can say, "Well, this clustering technique was better than that one because it corresponded more to the way that we use the decision."

GALISON: It's easier if you have other independent tests and you could, say, go to a higher frequency.

HILLIS: There's a great example of that, which was done by accident with the example you said about the type specimens. The grouping of animals, in terms of genus, and species, and things like that, was done by people deciding that the important character is the shape of the jaw, or the number of tailbones, or something like that. What was interesting was that was all done pre-DNA and even pre-necessarily everybody that was doing it believing in evolution. But when we got the ability to sequence mitochondrial DNA and got some insight into the process, it turns out almost exactly all those judgments were correct. They had picked the correct character and so on, so their clustering method, although it seemed very arbitrary, in fact, exactly reproduced the tree of life.

GALISON: If you're an explorer hiking in the Amazon and you find this turtle that you think is a new turtle—no one has seen a turtle that's of this species—you would probably not choose one that looked almost like the extreme example of the new turtle that looked a lot like an old turtle; you'd look for a turtle that was pretty different. So, you're prejudiced towards a type specimen that is more distinct than the marginal one might be.

MCEWAN: I'm thinking here of Leeuwenhoek's submissions to the Royal Society, where he is drawing a sperm and he inserts a homunculus.

GALISON: That seems to me an early example of what I call the spring of Narcissus problem. No one knows what a supermassive black hole, what the form of the shadow is going to be. There are models and simulations, but no one knows. It's not like looking at a known galaxy and then saying, "How does my method match up with what we've already seen at not as good gradient telescopes."

GERSHENFELD: Something that is misleading in the way we've been talking about this is modern clustering algorithms don't give you true force; they give you distributions. A hard clustering like k-means will miss something important just over the boundary. It's probabilistic. Applied to this, you don't get the image, you get PDFs over families of images. That's how modern clusters work. In modern classification, you don't get the classification, you get the probabilities of associations, and the classifier wouldn't tell you the difference between 49, 51 and 0, 100 likelihood.

Modern classifiers give you the classification, but they also give you uncertainty on top of that.

GOPNIK: There's an interesting contrast if you're looking at humans, and especially if you're looking at kids, in that one of the things that people have discovered that's interesting is if you're looking at kids categorizations, there certainly seemed to be some kinds of processes that are doing things in an associative way that are essentially looking for distributions. But by the time kids are linguistically categorizing things, they have something that is a much more essentialist, science-like category, a natural kind category. So, what they think is the thing that you're pointing to when you say something about a dog, it has nothing to do with any distribution of the properties of a dog. They think, it's whatever is the underlying causal category, whatever is the underlying causal set of properties, which is giving rise to some set of data, some set of things that you're perceiving, which could turn out to be completely wrong.

That abstract notion about what a category is that comes in science, the natural kind idea that it's whatever is out there in the world that's causing this set of correlations among data, that seems to be what the four-year-olds think that a category is and not the idea of the distribution. You can show that they can detach the distribution from the time they're infants, so they are in effect doing the clustering, but their conception of what's going on with the clustering, even when you're three or four years old, is that it's this abstract underlying causal system that's giving rise. The Bayesian picture seems to be very deeply built into the way we were even thinking about categories.

GALISON: It may be in the trees that stand around us in understanding that AI will come in as more than one tree and that there may be different ways that AI will function in that. It won't be just the AI tree, and the differential equation tree, and the analog model tree, and so on; rather, AI may stand in different ways and in different forms of clustering, in particulars and probabilistic distributions and so on. We ought to remain open to that possibility, too, that more than one AI in the image context will help constitute what we mean by understanding in ten years.

SETH LLOYD
Communal Intelligence

We haven't talked about the socialization of intelligence very much. We talked a lot about intelligence as being individual human things, yet the thing that distinguishes humans from other animals is our possession of human language, which allows us both to think and communicate in ways that other animals don't appear to be able to. This gives us a cooperative power as a global organism, which is causing lots of trouble. If I were another species, I'd be pretty damn pissed off right now. What makes human beings effective is not their individual intelligences, though there are many very intelligent people in this room, but their communal intelligence.

SETH LLOYD is a theoretical physicist at MIT; Nam P. Suh Professor in the Department of Mechanical Engineering; external professor at the Santa Fe Institute; and author of *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos.*

* * * *

SETH LLOYD: I'm a bit embarrassed because I've benefited so much by going close to last in this meeting. I've heard so many wonderful things and so many great ideas, which I will shamelessly parrot while trying to ascribe them to the people who mentioned them. This has been a fantastic meeting.

When John first talked about doing something like the Macy Conferences, I didn't know what they were, so I went back and started to look at that. It was remarkable how prescient the ideas seemed to be. I couldn't understand that, because why was it that all of a sudden we're now extremely worried and interested in AI and devices that mimic neural networks? People were worried about it back then, and yet for decades it didn't seem like people were that worried about this.

Rod Brooks made the point that what happened was the digital revolution took off. Moore's law went ahead full steam, and anything that wasn't a von Neumann architecture just wasn't worth doing because you would soon have a von Neumann machine that would be able to do anything that you could do. People just rather stopped worrying about this for a while.

Now, however, we're in quite a different era. I do have some things I'd like to say about artificial intelligence and even about quantum machine learning, but I'd like to give a little perspective about Moore's law. This is from someone who's trying to build computers where you store bits of

information on individual atoms and on superconducting quantum computers, and also with people who are trying to extend Moore's law further and further.

We're not at the end of Moore's law right now, but various aspects of it ended long ago. Most noticeably, the processor speed, which had been doubling every few years, crapped out at about three gigahertz around fifteen years ago—around 2003 or something like that—simply so the devices wouldn't melt. This led to the development of multi-core systems, which are primitive parallelism compared with Danny Hillis's connection machine but, nonetheless, a form of parallelism.

Now, as people are trying to press down to make the field effect transistor smaller and smaller, quantum mechanical tunneling effects are coming into play, and leakage current is growing when you start to make these transistors smaller than five nanometers or so. At that scale, statistical fluctuations in the number of electrons on the transistor comes into play, and the amount of noise that's going in the system grows, the wiring problem gets worse. It's clear that you can't just have more of the same of Moore's law. Just making von Neumann-like Intel processors is not going to keep going for that much longer.

What's happening is not that Moore's law is ending, but it's fragmenting into a variety of different kinds of systems. People are already using GPUs to do lots of these neural network systems. Field-programmable gate arrays are extremely useful for fast control systems. Neuromorphic computation is being explored, where you make systems that are more analog.

I have to say a little bit about analog versus digital here even though it's a false dichotomy. When John said he's going to make us all vote for analog or digital, Danny said, "But that's so digital of you." The system's bottom nature is quantum mechanical, as Freeman Dyson pointed out, and quantum mechanics is both analog and digital. Once you operate at this very small scale, the digital nature of the universe is extremely important.

The kind of information processing that Caroline Jones was talking about, information processing that's going on in the gut, suggests a new set of apps to enlist your gut to compute for you or to enlist your gut to give you the gut feeling of whether this is a spiral galaxy or an elliptical galaxy.

It's important to note, and Neil Gershenfeld pointed this out, that by far the largest amount of information processing going on in the human body is not in the brain; it's digital-chemical information processing that's going on at

the level of DNA and RNA, which is the ultimate digital forum for information, because quantum mechanics makes nature digital. It gives you only a certain number of types of elementary particles, which are combined to make only a certain number of types of atoms, which combine to make a large but countable number of molecules. They can be in different places. Somewhere, billions of years ago, living systems figured out how to harness this very microscopic digital nature of nature into encoding genetic information into DNA and RNA, and into the receptor dynamics and the receptors in cells. All cells have receptor dynamics in the metabolism of cells.

As Neil pointed out, if you look at what's going on in the genetic reproduction in a cell, it takes about a second to bring in something, but there are 10^{18} operations per second. Whereas, the brain has roughly 10^{11} neurons, 10^{15} synapses, and is going at 100 hertz—that's only 10^{17} operations per second. These are very large numbers. This has been going on for billions of years. Neurons haven't been around for billions of years, but by god cells have been, and they've been processing information very effectively in a way that combines analog and digital methods.

A wonderful insight for what happened came from Frank Wilczek's talk. I agree that there is no singularity that's going to be taking place anytime soon. Moreover, it is a pity that there aren't more West Coast people here, because when I go out there I find that a large number of Silicon Valley billionaires seem to believe that the singularity is there and that they themselves will be uploading their consciousness into a computer sometime in the near future.

Moreover, John was talking about what happens if you don't read some well-known books. I suspect that if you uploaded yourself to the cloud, even if it were entirely successful and you found yourself as yourself in the cloud but unable to go out for a cappuccino, you might feel that you'd struck a Faustian bargain by definition. There are plenty of stories about people who desire to live forever and the technologies they use. I don't ever remember any one that worked out well, unless maybe you count the New Testament, and I'm not sure we should count that.

ALISON GOPNIK: I had a conversation with a young man at Google at one point who was very keen on the singularity, and I said, "One of the ways that we achieve immortality is by having close relationships with other people—by getting married, by having children." He said that was too much trouble, even having a girlfriend. He'd much rather upload himself into the cloud than actually have a girlfriend. That was a much easier process.

LLOYD: This reminds me of my course at MIT. I write the problems on the board (they're not posted online). If you want the problems, you either have to go to class or you have to make a friend. I said, "For you MIT students, you'll have to decide which is harder to do." Let me just say that class attendance is very good.

My mother just died. It was very sad, and I'm still trying to understand that. Of course that's the kind of immortality that's worth going for and not the immortality of writing wonderful books or doing great science, even though that's also a good kind of immortality to strive for. As you say, what's important are the parts of yourself that you leave with the ones whom you love and whom are important to you that propagate in good ways.

This is what I loved about what Frank was saying. If you just look at these numbers for building new devices—and we are going to be building beautiful, huge new devices that have vast amounts of information processing power—that, in the not-so-distant future, will match this roughly 10^{17} ops per second on something like 10^{15} bits. That's something that is likely to happen in the next half century or so, though it's not going to be by a von Neumann architecture. It's going to have to be by a variety of different methods.

As discussed by David Chalmers in his talks about consciousness, and emphasized by Rod and Danny and others, people already treat the artificial intelligences in their life as very important companions that they would never be without. By becoming accustomed to treating these artificial intelligences as though they're alive, even if it might not meet the criteria for being able to perceive a gestault is, as I mentioned before, one of the main issues that was brought up back in the Macy Conferences, the early ones. Can an artificial intelligence have a gestault?

Even if we have something that we know for sure is not conscious, doesn't have gestalt, and it's a very simple circuit, we still feel for it and don't want to cause it pain. We haven't talked about the socialization of intelligence very much. We talked a lot about intelligence as being individual human things, yet the thing that distinguishes humans from other animals is our possession of human language, which allows us both to think and communicate in ways that other animals don't appear to be able to. This gives us a cooperative power as a global organism, which is causing lots of trouble. If I were another species, I'd be pretty damn pissed off right now. What makes human beings effective is not their individual intelligences, though there are many very intelligent people in this room, but their communal intelligence.

My prediction would be that there's not going to a singularity. But we are going to have devices that are more and more intelligent. We'll gradually incorporate them in our lives. We already are. And we will learn about ways to help each other. I suspect that this is going to be pretty good. It's already the case that when new information processing technologies are developed, you can start using your mind for different things. When writing was developed—the original digital technology—that put Homer and other people who memorized gigantic long poems out of a job. When printing was developed and texts were widely available, people complained that the skills they had for memorizing large amounts of things and poetry—which is still a wonderful thing to do—deteriorated.

There's plenty of evidence that the way people use their memory, given that they have immediate access to Internet search, changes a lot. For myself, I'll just say that I no longer remember what it was, I just remember what I did to get it. Where did I go? What were the search terms I used to find this? Then I can find it again. Let's not even mention the fact that nobody knows where the heck they're going in their head any longer because they just have somebody saying, "Turn left at the next intersection."

This is going to be very interesting. If we think of artificial intelligence as part of the human communal development, then this is going to be very empowering for us and for these artificial intelligences. There are a lot of bad things out there. The fact that the largest amounts of artificial intelligence out there are being used by large corporations to sell us crap we don't need, I sometimes question their intelligence. I've had both my hips replaced, and I frequently get these ads saying, "Dear Seth, you have this artificial hip. Perhaps you'd like to try this other one. Oh, and by the way, here's a Swiss army knife for you to do it yourself." What are they thinking? I don't get it.

Moreover, the question is what they could do with that information should they choose. If Google were more like the government of China or if Google reenters China and the government of China asks it to do things for the government of China, then we are in something that's much worse than *1984* at some level. That's stuff to worry about. This notion was popular with Stephen Hawking and Elon Musk, that we'll create a maligned artificial intelligence that will take over society. It just seems silly. First of all, we're far away from having such an artificial intelligence. We'll have, I would say, centuries before such a thing might exist, and we have plenty of time to make sure that if such a thing exists that we'll be okay.

Reading is helpful for this. We know that if you create an artificial being who is both more intelligent, stronger, and more ethical than you, as Mary

Shelley pointed out, you better not treat it as if it's subhuman. If you do, then it will behave in a psychotic fashion. If we simply choose to be kind to the artificial intelligences that we create, we'll be going a long way in the right direction. We should also be very careful about the companies that are spying on us and are using artificial intelligence primarily to sell us useless crap over the Internet.

Amongst these technologies that are likely to be useful, these novel technologies of information processing, are quantum computers, which have not yet done anything that a classical computer couldn't do. However, despite the fact that they're still piddling and tiny, they now have fifty quantum bits and hundreds of thousands of quantum bit quantum computers are likely to show up soon. These are going to be just one of these information processing tools. They're now at the stage where they can process information for specialized problems like simulating other physical systems, an application proposed by Richard Feynman, that they can do better than classical supercomputers. That's going to keep on going.

About six or seven years ago, my post docs and I began looking at applying quantum information processing to do machine learning. The simple intuition is that quantum systems can generate statistics that cannot be generated by any classical computer equipped with a random number generator. They can generate strange and counterintuitive phenomenon. This has been known for more than a century. We also know from the example of things like deep neural networks, or Boltzmann machines, or deep learning that if you build a device that can generate certain kinds of statistics, it can often be used to recognize similar kinds of patterns. So, if quantum systems can generate patterns that cannot be generated classically, perhaps they can also recognize and categorize patterns that can't be categorized or recognized by a classical system. Moreover, these might go beyond what weirdness like the EPR effect and stuff like that. It might also be that they can find patterns in nature for things that you could never do on a classical computer.

For example, what we first started out doing is exactly these k-means, quantum k-means, and quantum support vector machines, and then moving on to just bread and butter things like regression and principle component analysis, matrix completion (the Netflix algorithm). These are methods that involve linear algebra, and a lot of learning techniques just involve taking gigantic vectors of data and multiplying them by humongous matrices and applying some kind of nonlinear transformation, and then you do it again and you try to train the system to work. Well, quantum mechanics is about humongous vectors and gigantic vector spaces and multiplying them by gigantic matrices, and then doing nonlinear things like measuring and

then seeing what happens. If you do encode data in a quantum mechanical state, you can kick serious machine-learning ass. Even with Google's 50-qubit superconducting quantum computer, you could in principle diagonalize a 10^{12} by 10^{12} matrix, something which would take Avogadro's number of operations ordinarily, and you're not going to do that classically for quite a while.

* * * *

W. DANIEL HILLIS: You touched on something that I went back and read because you had mentioned it in an earlier conversation. In the early Macy Conference, in Ashby's discussion on the chess-playing computer, he talks about an algorithmic chess player, but in his formulation, besides a general purpose machine, he also includes a Geiger counter. He seems to think somehow that this is important. Going back to Alison's point, Bigelow says, "I agree, it's different with that, but why don't we just throw that away, and it'll all work just as well." Which is in fact what happened, and that was the truth. They were correct that the machine with a true element of randomness was different than a classical machine; it just wasn't different in the way that was helpful.

LLOYD: That's an interesting point. Since you mentioned that, I also thought about that some more, about where randomness plays a role. Well, neurons and synapses are noisy because there are small numbers of chemicals. So, neural functioning is quite noisy. The kind of digital cellular level information processing in terms of genetic reproduction is very precise. Nine out of ten of the offspring of an E. coli have exactly the same DNA as the original E. coli, but of course we know that it's useful to have stochastic processes. In fact, if you stress the E. coli by putting in a bit of alcohol or something in their petri dish, then they start making more mistakes because they're in a bad genetic place.

This is related to what Neil was saying about state of the art machine learning algorithms. In game theory, what is Nash equilibrium? Nash's beautiful theorem says that if you have a game, then there are these equilibria where both players can't change what they're doing without making things worse for themselves. But in order to achieve that, you need a probabilistic strategy. In order to apply the Kakutani fixed-point theorem, you need a continuous space of strategies so that you could say, "If I change my strategy, it's not going to work." The best strategies then are these probabilistic strategies. Plenty of times this is a very good thing to do.

HILLIS: But it doesn't require true randomness. Pseudorandomness works just fine.

FRANK WILCZEK: Although, there have been scientific applications where pseudorandom numbers ran into trouble.

LLOYD: Right. Pseudorandomness can be problematic. It's expensive computationally and, by definition, it is not random. So, if you happen to hit one of those non-randomnesses at the wrong time, it could cause you trouble.

NEIL GERSHENFELD: What's your take on the power of partially coherent quantum computers? So, quantum computers, the real true ones are maximally coherent, which means they can be completely entangled, and a lot of the things called quantum computers that have huge numbers of bits are only a little bit coherent, and there's a big debate about how useful they are.

LLOYD: D-Wave is not a full-blown quantum computer; it's a quantum annealer. You encode the answer to a hard problem in the ground state of a system. If you can find the lowest energy state, then you've solved the problem, which is a classical method for doing this as well. As a result, they're much more immune to noise, the fact that they're rather incoherent.

The lowest state is the answer. There's a classical form of this called simulated annealing, where you set up the logical constraints of your problem so that the energy is the number of violated logical constraints. So, the ground state by definition has the lowest energy because none of the constraints are violated. So, it's a solution. And then you cool it down to try to find the answer.

GERSHENFELD: Another way to say it is you put it in the answer, but you change the question. If you put it in the answer to an easy problem, you then deform it to asking a hard problem, and if you change it slowly enough it stays in the answer.

LLOYD: Quantum annealing is based on what Neil just said: You start at a very easy thing to say, like all the spins in your computer should be pointing this way, and then you gradually turn on this energy function that you wish to find the lowest energy state. There's a theorem called the adiabatic theorem that says if you do this slowly enough you'll get there.

This notion of doing computation this way, quantum computation the way it was developed at MIT, but the design for the D-Wave system was developed by my graduate student, Bill Kaminsky and me in 2002. We failed to patent it because we did a little calculation, and we said, "Well, after you've entangled about 50 quantum bits, then even under the absolute most optimistic assumptions, that is not going to work. The energy will be too high." Then D-Wave spent $100 billion building this from which I conclude that you should always patent things even if you're absolutely sure that they're not going to work.

The D-Wave system is partially coherent. It does solve hard problems. In fact, you can show that having a bunch of noise in the middle is helpful for it. It can very well be helpful for it to have noise in the middle. There are plenty of kinds of computation, including things that were developed by Shannon and von Neumann's stochastic computing, which were not adopted. They were developed back in the '40s and '50s but not adopted because of the power of rapidly increasing power of digital computers.

Once you start pressing Moore's law, your systems are going to be noisy. They are going to be stochastic. They're going to be quantum mechanical, but they're going to be semiquantum mechanical. They're going to be semicoherent. This is a wonderful opportunity to develop a theory and practice of these kinds of computers, which will be the most powerful computers that you could build, where you have to deal with noise and you have to deal with quantum mechanics.

DAVID CHALMERS: The point at which machines achieve human level capacities in a wide range of areas, one of the areas where they'll be at human level capacity is creating artificial intelligences. The moment they get a little bit beyond human level capacities, they'll be a little bit beyond human level capacities at creating AI, therefore they'll be able to create AI systems a bit better than those that we can create. Therefore, they'll be able to create AI systems a bit better than themselves. Iterate until superintelligence. That's always struck me as a very promising argument. Do you think there's something wrong with that?

WILCZEK: Things can increase and saturate a bound, or they can take off, or they can do something. They can slowly increase. There's nothing inevitable about a singularity. The structure of high problems, P versus NP, suggests that there are going to be problems where progress will be very slow.

CHALMERS: Why does it have to be inevitable to be interesting? This happens a lot in arguments about this. You don't know that's going to happen. Even if there's a 10 percent chance it's going to happen, that's interesting.

HILLIS: There's a flaw in the description, which is that it suggests that intelligence is this uni-dimensional thing. Something can be incredibly smart and not have the ability to make a remotely smart machine. You're assuming a particular dimension of intelligence could go off in that direction, but it would be a very narrow dimension.

CHALMERS: Once you have correlations between capacities, if one dimension goes off, then the things that correlate with it will tend to go off. If one of the things which goes off to infinity is the ability to create AI, then at the very least we get this offshoot line.

LLOYD: First of all, can we just do some numbers again? It's not going to go off to infinity. Computation is a physical process, indeed, as a number of people in this room are fond of claiming that all of physical dynamics can be thought of as a computation, as information processing, and there's only a certain amount of information processing you can do. Now, those amounts are large. If you're willing to turn things into black hole density and compute using black holes or something, but that's unlikely to happen. If you say we're going to compute using things that have electrons and ordinary materials that are held together by covalent bonds, then you're going to have basically ops operating at the level of an electron vault or something like that, and that's where nature is doing it already.

GOPNIK: It's curious because if you think about it, we already do that. We do know that the current intelligence that we have, one of its characteristics is that it creates intelligences that are superior to it on a regular basis, which in turn create intelligences that are superior to those intelligences. It doesn't seem to bother us very much, presumably because we die before we get to great grandchildren, but that process is taking place. It doesn't strike anyone as being particularly maligned that we're creating generations that are capable of doing things that we're not capable of doing.

CHALMERS: Every PhD advisor is trying to create an intelligence greater than theirs.

GOPNIK: In fact, literally succeeding. Right? That's the whole plan of how human intelligence works, and it is interesting that it strikes us as being hopeful rather than striking us as being maligned.

GERSHENFELD: I find the problem to be Ray Kurzweil's followers, not him. A lot of what Ray does is he projects data. If you look at this data, Ray himself does a good job, and if you just look at the data he projects, it's an interesting moment. The data projects in an interesting way. It's about singularity, but do look at Ray's data. The data is interesting.

LLOYD: There has been this old projection. It's been noted for at least fifty years that human population is growing super exponentially. As the rate of growth of the population goes, of course it's proportional to the number of people there, but there's another positive term that's proportional to the square of the number of people, which is the number of possible interactions you can have.

The way I make sense of this is exactly because we do have this funky universal human language, and because our intelligence is a communal intelligence, that our capacity comes from not just how many people there are, it's how many interactions there are between people, and this gives you this proportion of the square. If you integrate that, you find that the population becomes infinite, and if you extrapolate from historical amounts of population, it becomes infinite at something like 2070. It becomes infinite in half a century or something like that. Luckily, it slowed down recently. There are these trends toward singularity.

CAROLINE JONES: People get stupider, too. On the many axes of intelligence, there are many axes right now where people are extinctifying themselves. That's stupid. That's a massive failure of intelligence.

LLOYD: We overemphasize. As artificial intelligences get closer to the capacities of human beings, they are already exhibiting behaviors that are very human-like, messing up in weird and inscrutable ways that we don't understand. Artificial intelligence often leads to real stupidity, and that's one of the signs that it's intelligent. Human beings operate in a self-contradictory fashion. We don't do things rationally, and by god we shouldn't do things rationally, as you're arguing. Computers are going to do that as well. Deep neural networks are already being to design the next generation of programming systems. This is not some science fiction. This is happening already.

RODNEY BROOKS: Programming?

LLOYD: Maybe there's this distinction that's come up a bunch of times about what's the difference between a neural net that's been trained and a program that's been written into memory.

CHALMERS: I remember back in 1978 when I was a computer hobbyist at twelve years old, there was a program that was released called "The Last One," and it was going to be the program that wrote programs. Once you got the program to write programs, we're never going to need another one. It didn't quite work out.

STEPHEN WOLFRAM: So, as you realize the main problem is you have to specify what the thing is going to do. With respect to this question about ever increasing intelligence and so on, it will be nice to hear from people what they imagine the definition of intelligence from some physics mathematics point of view might be, because I think it's all nonsense. In the end you'll realize that intelligence is just computation, and you realize that computation happens in lots of kinds of systems. It happens in lots of systems in the universe. It's something where you say we're going to have this ever-increasing intelligence. This doesn't make any sense. The universe is already computing in a very efficient, effective way in all kinds of different places. The question is whether this computation is aligned with something that we think of as being human-like intelligent behavior, and that's a completely different question and one that is quite separate from all these singularity discussions.

CHALMERS: The cash value is doing things that we care about. Right? Like solving problems, curing diseases, winning wars.

LLOYD: That's very good point. As you know, Steve and I have both written books claiming the universe is a giant computer and that we should understand everything in terms of computation. What's going on is when we're building computers, particularly when we're building quantum computers, we're hacking into the ongoing computation that's going on and having more of that be computation that we'd like to have.

The real issues are not about the use of comp flops but about the use of joules and about energy that we're using. Those are the really hard ones. Then it's going to be okay. If we pay attention to the computers we're building, if we socialize them, we treat them nicely, they then are part of our human intelligence and not separate from it in the same way that books are not separate from our intelligence.

ROBERT AXELROD: I'm going to take your example of advertisements for hip replacement, which you labeled as stupid, and give an account of why it's intelligent. You know a lot more people that have had or will have hip replacements or are on the verge of having them than I do. You are a social collector of people who are relevant to hip advertisers. Even though you

won’t have need of one, you might find that the one advertised is better than the one you got.

JONES: But he doesn’t want to be a node in capitalism’s purchasing customers.

AXELROD: I’m just saying the capitalist system that’s advertising hips to him is not stupid. Where’s the intelligence that discovers that you’re a hip replacement node? The answer might be that it’s an automated system already that tests a lot of different ways of focusing ads and finds that people that have purchased something should still be advertised for the same thing, even though, as in your case, you know you’re not going to need another one. The system might have discovered that without anybody designing it to discover that, because they try a whole bunch of stuff and some of it gets good feedback in terms of selling hips or cars or whatever it is. So, it’s a combination. In this case, the intelligence could be accounted for as you’re doing some of the work by collecting hip-relevant people and talking to them when you learn something about hips. The advertising system is also learning that that works, so it’s a combination of human social intelligence and the automated system. It’s a good example we’ve been talking about of how those are going to merge and complement each other.

WILCZEK: It’s poetic that we’re close to the end and bringing together so many themes in terms of hip replacement, but it does illustrate opacity. It illustrates looking at extreme cases.

JOHN BROCKMAN: It gets better. The reason I was energized to do this project was because I went to get a cortisone shot, nothing major, but it was for a pain in my neck, which means they have to do it in a hospital setting. So, I make an appointment at the Hospital for Special Surgery at 3pm, get a cup of coffee, come back, hit my e-mail. First email: New England Burial Society. I get a second e-mail: New England Crematorium dot com. Third email: Casket dot com: "Keep your remains intact for a thousand years." This is very sophisticated because I knew that something was happening, and that something had to be deep learning. I immediately thought of Demis because I know this is beyond Larry Page. Why? Because I made the appointment from my farm in Connecticut, and who knew that I don’t do the boroughs? So, I’m not going to the Brooklyn crematorium. Because that’s where they are. They’re in the Bronx. They’re in Brooklyn.

WILCZEK: But it also illustrates what’s lacking. So, it has opacity. It has looking at extreme cases. What it doesn’t have is ...

LLOYD: Tact.

WILCZEK: It doesn’t have a sense of decency. That’s what we need is somehow to widen the circle of empathy on both sides.

LLOYD: Tact comes from the word to be silent. It’s something we could use. Herb Simon said the world that is information-rich is by necessity attention-poor. He said this in 1956 or something like that. That anticipated our current era. What we need to do as human beings is to protect our time and our attention, to pay attention to the things that are important such as other human beings and the odd, sexy AI.

BROCKMAN: Catherine Bateson asked, "Why can’t we have an AI with humility?" Why can’t we have an AI that asks the question and then says, "Maybe I better sleep on it"?

W. DANIEL HILLIS
Emergences

My perspective is closest to George Dyson's. I liked his introducing himself as being interested in intelligence in the wild. I will copy George in that. That is what I'm interested in, too, but it's with a perspective that makes it all in the wild. My interest in AI comes from a broader interest in a much more interesting question to which I have no answers (and can barely articulate the question): How do lots of simple things interacting emerge into something more complicated? Then how does that create the next system out of which that happens, and so on?

Consider the phenomenon, for instance, of chemicals organizing themselves into life, or single-cell organisms organizing themselves into multi-cellular organisms, or individual people organizing themselves into a society with language and things like that—I suspect that there's more of that organization to happen. The AI that I'm interested in is a higher level of that and, like George, I suspect that not only will it happen, but it probably already is happening, and we're going to have a lot of trouble perceiving it as it happens. We have trouble perceiving it because of this notion, which Ian McEwan so beautifully described, of the Golem being such a compelling idea that we get distracted by it, and we imagine it to be like that. That blinds us to being able to see it as it really is emerging. Not that I think such things are impossible, but I don't think those are going to be the first to emerge.

There's a pattern in all of those emergences, which is that they start out as analog systems of interaction, and then somehow—chemicals have chains of circular pathways that metabolize stuff from the outside world and turn into circular pathways that are metabolizing—what always happens going up to the next level is those analog systems invent a digital system, like DNA, where they start to abstract out the information processing. So, they put the information processing in a separate system of its own. From then on, the interesting story becomes the story in the information processing. The complexity happens more in the information processing system. That

certainly happens again with multi-cellular organisms. The information processing system is neurons, and they eventually go from just a bunch of cells to having this special information processing system, and that's where the action is in the brains and behavior. It drags along and makes much more complicated bodies much more interesting once you have behavior.

W. DANIEL HILLIS is an inventor, entrepreneur, and computer scientist, Judge Widney Professor of Engineering and Medicine at USC, and author of *The Pattern on the Stone: The Simple Ideas That Make Computers Work.*

* * * *

W. DANIEL HILLIS: My perspective is closest to George Dyson's. I liked his introducing himself as being interested in intelligence in the wild. I will copy George in that. That is what I'm interested in, too, but it's with a perspective that makes it all in the wild. My interest in AI comes from a broader interest in a much more interesting question to which I have no answers (and can barely articulate the question): How do lots of simple things interacting emerge into something more complicated? Then how does that create the next system out of which that happens, and so on?

Consider the phenomenon, for instance, of chemicals organizing themselves into life, or single-cell organisms organizing themselves into multi-cellular organisms, or individual people organizing themselves into a society with language and things like that—I suspect that there's more of that organization to happen. The AI that I'm interested in is a higher level of that and, like George, I suspect that not only will it happen, but it probably already is happening, and we're going to have a lot of trouble perceiving it as it happens. We have trouble perceiving it because of this notion, which Ian McEwan qso beautifully described, of the Golem being such a compelling idea that we get distracted by it, and we imagine it to be like that. That blinds us to being able to see it as it really is emerging. Not that I think such things are impossible, but I don't think those are going to be the first to emerge.

There's a pattern in all of those emergences, which is that they start out as analog systems of interaction, and then somehow—chemicals have chains of circular pathways that metabolize stuff from the outside world and turn into

circular pathways that are metabolizing—what always happens going up to the next level is those analog systems invent a digital system, like DNA, where they start to abstract out the information processing. So, they put the information processing in a separate system of its own. From then on, the interesting story becomes the story in the information processing. The complexity happens more in the information processing system. That certainly happens again with multi-cellular organisms. The information processing system is neurons, and they eventually go from just a bunch of cells to having this special information processing system, and that's where the action is in the brains and behavior. It drags along and makes much more complicated bodies much more interesting once you have behavior.

Of course, it makes humans much more interesting when they invent language and can start talking, but that's a way of externalizing the information processing. Writing is our form of DNA for culture, in some sense; it's this digital form that we invent for encoding knowledge. Then we start building machinery to do information processing, systems, everything from legal systems to communication systems and computers and things like that. I see that as a repeat pattern. I wish I could say that more precisely, but you all know what I'm talking about when I wave my hands in that direction. Somebody will someday make wonderful progress in finding a way of talking about that more precisely.

There's a worry that somehow artificial intelligence will become superpowerful and develop goals of its own that aren't the same as ours. One thing that I'd like to convince you of is that I believe that's starting to happen already. We do have intelligences that are superpowerful in some senses, not in every way, but in some dimensions they are much more powerful than we are, and in other dimensions much weaker. The interesting thing about them is that they are already developing emergent goals of their own that are not necessarily well aligned with our goals, with the goals of the people who created them, with the goals of the people they influence, with the goals of the people who feed them and sustain them, goals of the people who own them.

Those early intelligences are probably not conscious. It may be that there's one lurking inside Google or something. I can't perceive that. Corporations are examples. Nation states are examples. Corporations are artificial bodies.

That's what the word means. They're artificial entities that are constructed to serve us, but in fact what happens is that they don't end up serving exactly the founders, or the shareholders, not the employees that they serve, or their customers. They have a life of their own. In fact, none of those entities that are the constituents have control over them. There's a very fundamental reason why they don't. It's Ashby's Law of Requisite Variety, which states that in order to control something, you have to have as many states as the thing you're controlling. Therefore, these supercomplicated superintelligences, by definition, are not controllable by individuals.

Certainly, you might imagine that the head of Google gets to decide what Google does, especially since they're the founder of Google, but when you talk to heads of state or things like that, they constantly express frustration that people imagine that they can solve this problem. Of course, shareholders try to influence and do influence corporations, but they have limited influence.

One of the interesting things about the emergence of them having goals of their own is the emergent goals often tend to successfully see those influences as sources of noise, or something like that. For example, before information technology, corporations couldn't get very big because they just couldn't hold together.

BROOKS: What about the East India Company?

AXELROD: Or China.

HILLIS: I would say that East India Company did not as effectively hold together as an entity and stay coordinated. They can be big, but I don't think that they were as tightly coupled.

Information technology certainly made it much easier. I won't quibble with you whether they were edge cases, but you could have skyscrapers full of people that did nothing but hold the corporation together by calling up other people in the corporation.

These things are hybrids of technology and people. As they transitioned to a point where more decisions were being made by the technology, one thing they could do was prevent the people from breaking the rules. It used to be that an individual employee could just decide not to apply the company policy because it didn't make sense, or it wasn't kind, or something like that. That's getting harder and harder to do because more of the machines have the policy coded into it, and they literally can't solve your problem even if they want to.

We've got to the point where we do have these superpowerful things that do have big influences on our lives, and they're interacting with each other. Facebook is a great example. There's an emergent property of Facebook enabling conspiracy theory groups. It wasn't that Zuckerberg decided to do that or anybody at Facebook decided to do that, but it emerged out of what their business model was. Then that had an impact on this other emergent thing—the government—which was designed for dealing with people, not corporations. But in fact, corporations have learned to hack it, and they've learned that they can use their superhuman abilities to track details to things like lobbying and track details of bills going through Congress in ways that no individual can. They can influence government in ways that individuals can't. More and more, government is responding to the pressures of corporations more successfully than to the pressures of people because they're superhuman in their ability to do that, even though they may be very dumb in some other ways.

One of their successes is their ability to gather resources; to get food from the outside world, for example. They have been extremely successful at gathering resources to themselves, which gives them more power. There's a positive feedback loop there, which lets them invest in quantum computers and AI, which gets them presumably richer and better.

We may be already in a world where we have this runaway situation, which is not necessarily aligned with our individual human goals. People are perceiving aspects of it, but I don't think what's happening is widely perceived. What's happening is that we have these emergent intelligences. When I hear people do this hypothetical handwringing about these superintelligent AIs that are going to take over the world, well, that might happen some time in the future, but we have a real example now.

Why don't we just figure out how to control those, rather than thinking hypothetically how we ought to design the five laws of robotics into these hypothetical general AI human-like things? Let's think how we can design the five laws of robotics or computers into corporations or something like that. That ought to be an easier job. If we could do that, we ought to be able to apply that right now.

* * * *

ROBERT AXELROD: An example of that is, what rights do they have? The Supreme Court recently said they had the right to free speech, which means they can contribute to political campaigns.

ALISON GOPNIK: David Runciman, who is a historian at Cambridge, has made this argument exactly about corporations and nation states, but he's made the argument—which I think is quite convincing—that this is from the origin of corporations and nation states, that it's from industrialization, that that's when you start getting these agents.

Then there are some questions you could ask about whether you had analogous superindividual agents early on. Maybe just having a forager community is already having a superintelligence, compared to the individual member community. It's fairly clear that that kind of increased social complexity is deeply related to some of the things that we more typically think of as being intelligences. We have a historical example of those things appearing and those things changing the way that human beings function in important and significant ways.

For what it's worth, at the same time, the data is that individual human goals got much better on average. You could certainly argue that there were things that happened with industrialization that set back.

AXELROD: What do you mean goals got better?

GOPNIK: Well, people got healthier.

AXELROD: They achieved their goals.

GOPNIK: Yes, exactly. They stopped having accidents. They stopped being struck by lightning. Someone like Hans Rosling has these long lists that are like that. We do have a historical example of these superhuman intelligences happening, and it could have been that people thought the effect was going to be that individual goals would be frustrated. If you were trying to graze your sheep on the commons, then you weren't better off as a result, but it certainly doesn't seem like there's any principle that says that what would happen is that the goals of the corporations would be misaligned.

W. DANIEL HILLIS: It's a matter of power balance. Certainly, humans aren't powerless to influence those goals. We may be moving toward tipping the balance, because a lot of technological things have helped enable the power of these very large corporations to coordinate, and act, and gather resources to themselves more than they've enabled the power of individuals to influence them.

RODNEY BROOKS: Back to the East India Company: I realized when I said that that in fact the East India Company did develop an information technology and became the education system through elementary schools of people being able to write uniformly, do calculations, arithmetic. Writing enabled their information technology that individual clerks were substitutable across their whole operation.

HILLIS: The East India Company did some pretty inhuman things.

NEIL GERSHENFELD: Al Gore said he viewed the Constitution as a program written for a distributed computer. It is a really interesting comment, that if you take what you're saying seriously to think about what is the programming language.

STEPHEN WOLFRAM: It's legalese. Programming language is legalese.

CAROLINE JONES: That the algorithms of homophily are a huge part of the problem. The reputed echo chamber that magnifies small differences so you get conspiracy theories—the schizophrenic model is hyper connectivity. Everything connects to this conspiracy theoretical model, so homophily, as I learned from Wendy Chun, is at its core of the programming language—like begets like—as distinguished from the parallel study in the '50s of birds of a

feather don't flock together; difference attracts. These were two models in the '50s that were at the core of this game theoretical algorithmic thinking, and everyone went with like begets like, which produces the echo chamber.

The first question is about hybridity. The DNA model has been radically complicated by translocation. So, it's not the case that there are perfect clones. You mentioned nine out of ten E. coli, but there's the one tenth, which has information from the chimeric gene that I have floating around me from my son when he was passing in my amniotic fluid, whatever. There's translocation going on all the time.

In other words, do we have a resource there in this ongoing hybridization of the program? Do we have a resource point of inflection? To Bob's rights comment, we also are giving rights, not "we," but the Bolivian constitution is giving rights to the ocean, to a tree, to cetaceans. So, can this dialogue with other life forms, with other sentiences somehow break the horrifying picture of the corporate superintelligence? Are there other translocatable informational streams that can be magnified or the algorithms be switched to proliferate differences and dialogue and external influences rather than the continuous proliferation of the self same?

HILLIS: I don't think it's necessarily horrifying, because I don't think we have *no* influence over this. I agree that this has been going on for a long time.

JONES: But we do have the model of a government being put in place by algorithms that we no longer control demographically. We have an actual case.

HILLIS: The trend is very much in the direction of the next level of organization, which is corporations, nation states, and things like that taking advantage of these effects, like symbiosis.

WOLFRAM: That's called strategic partnerships.

HILLIS: Exactly. Yes, it is, or acquisition of genetic material is done by acquisition. They have lots of ways of taking advantage of hybridization that is better than individuals. In fact, the technology has hurt the individual

interactions, as you point out, with the way that it's played out and, in many ways, harmed it. It's helped it in some ways.

It's been a mixed bag, but it's definitely enabled the corporations because corporations before were limited just by the logistics of scale. They became more and more inefficient except in very special cases. They couldn't hold together as they got bigger. Technology has given them the power to hold together and act effectively bigger and bigger, which is now why we've just gotten in the last year the first two trillion-dollar companies because they were designed from the beginning to take good advantage of technology.

PETER GALISON: Do you think that there's a characteristic difference between the kind of research that goes on under the corporate umbrella and, say, the university umbrella? I know people have lots of views about this, and there are things you can do in university that you can't do in one or the other, but how would you characterize in particular areas of AI-related work?

HILLIS: Corporations are much more rationally self-interested in how they focus their research.

AXELROD: You mean they're allocating resources more efficiently? They're more effective at promoting promising research areas? Is that what you're suggesting?

HILLIS: They select research areas that are in alignment with their emergent goals.

BROOKS: Yes, but they're doing an additional thing now, which is very interesting. They're taking the cream from the universities, offering them very open intellectual positions as a way of attracting the level below who will be more steerable to what they do. So, Google and Facebook are both doing this in the extreme at the moment. Those particular people will tell you what great freedom they have.

HILLIS: I'd say that's a great example of them being very smart and effective at channeling the energy toward their emergent goals.

WOLFRAM: As you look at the emergent goals of corporations, it's difficult to map how the goals of humans have evolved over the years, but I'm curious as to whether you can say anything about what you think the trend of emergent goals in corporations is. That is, if you talk about human goals, you can say something about how human goals have evolved over the last few thousand years. Some goals have remained the same. Some goals have changed.

AXELROD: I'll try my hand at it. When you get two corporations in the same niche that are competitive, they often become uncompetitive. If one of them is substantially bigger, they might try to destroy or gobble up the other one, but otherwise it might try to cooperate with the other one against the interest of the consumer. It's called anti-trust.

As they get bigger, they also want to control their broader environment like regulations. A small restaurant is not going to try to control the regulation of restaurants, but if you have a huge chain, then you can try to control the governmental context at which you are, and you could also try to control the consumer side of it, too. Advertising is a simple way to do that. As the corporations get bigger, there's an unfortunate tendency that the industrial competition goes down, and we see this in high tech. It's very extreme.

There are only five huge corporations and they're doing different things. Apple is doing manufacturing and Amazon is not doing much manufacturing. That's likely to continue not just in the high-tech areas, but in others. It's very worrisome that the corporations will get more and more resources to shape their own environment.

At the lower level—at a restaurant or something—you have two goals: make money for your owners and survive. But when you get much bigger it seems to me that often the goals beyond those two are to also control as much of your environment as you can.

WOLFRAM: For the purpose of stability or for further growth.

AXELROD: For both. There's another trend that's correlated with this, which is the concentration of capital. At the individual level, you see a higher and higher proportion of the wealth of a country is in the top one percent.

HILLIS: That's a symptom of them getting more powerful.

AXELROD: Maybe. It's a symptom of the returns on capital greater than the growth of productivity, which doesn't depend so much on the level of organizational structure. So, the corporations are likely to have more and more control over resources, and that's unfortunate. It's a very risky thing.

WOLFRAM: So, it's virtues and vices of corporations. Do you think the corporations will emerge with the same kinds of virtue and vice type goal structures that are attributed to humans?

GEORGE DYSON: One thing that is very much Danny's work, and that he didn't say, is that the world we inherited from the 1940s that brought the first Macy Conference, the huge competition was in faster computers, to break the code within 24 hours, to design the bombs. These were machines just trying to get more instructions per second.

But there's another side to it. There's slow computing that in the end holds the survival of the species, and that's where the immune system is so good because of very long-term memory, and we need that too. We don't just need the speed. Danny, of course, is building the 10,000-year clock, a very slow computer, and that's an important thing because when you have these larger organizations, these superorganizations you're talking about, they scale not only in size and distance but in time, and that's a good thing—or it can be a bad thing, too. You can have a dictator that lasts for a thousand years.

GOPNIK: But some organizations don't scale. Even when they get bigger, they seem to have this very predictable life. That's what people like Geoffrey West would say.

G. DYSON: Right. Geoffrey will say that. But a very important, possibly good, function of these systems is we're going to get longer-term computing where you look at the very long-term time series. That evolution will be a good thing.

GALISON: Historically, we have places like AT&T, IBM, Xerox that had world-class labs that deteriorated over time. AT&T Laboratories is nothing

remotely like what it was like in the 1960s and '50s, and they expelled a lot of research eventually because it wasn't short-term enough for them, and they figured they'd offload that to the universities and then take the fruits of it and do things that were more short term.

One possible outcome is that even the places where they're hiring people at a high level and giving a tranche of the research group relative freedom as a cover and attractor, one outcome is that that could expand, but it could also pull back, and you could end up with wrecking parts of the university and not having a lot of freedom in the corporation. I don't know. It seems to me an open question what's going to happen with this concentration of research wealth at a few companies.

BROOKS: The wealth is the important part. When AT&T labs was riding high, AT&T was a monopoly of the phone company, an incredible cash flow.

FRANK WILCZEK: They were required by law to spend money.

WOLFRAM: But the fact is, basic research happens when there's a monopoly, because if you have a monopoly then it's worth your while to do basic research because whatever is figured out will only benefit you. You see that even at the level of the U.S. government.

JONES: Did you hear Frank's comment that AT&T was required by the government to do research?

WILCZEK: They were required by law to keep their profits at a certain level, so they spent a lot on research.

JONES: A monopoly will never regulate itself.

WOLFRAM: Even in our tiny corners of the technology world, it's worth our while to do research in things where we are the only distribution channel basically, and the same thing is happening with a bunch of AI stuff that's being done in places where the only beneficiary is a company with a large distribution channel that there's motivation to do basic research there. As soon as you remove that monopoly, the motivation to do basic research goes away from a rational corporate point of view.

TOM GRIFFITHS: There are cases where you can tie this very directly to AI. The best example of this is the Facebook feed management algorithm. Nick Bostrom has this thought experiment where you make an AI whose goal is to manufacture paperclips, and then it consumes the entire earth manufacturing paperclips. Tristan Harris has pointed out that the Facebook feed management algorithm is essentially that machine, but for human attention. It consumes your attention. It makes money as a consequence of doing so that's fed back into the mechanism for consuming human attention. It gets better and better at consuming human attention until we've paper-clipped ourselves.

SETH LLOYD: That's true for all of these companies. Anybody who has teenage children knows that there's an attention problem.

GOPNIK: I would push back against that. That idea is highly exaggerated and let me give you the reason why I think that.

Think about walking or driving down a street where there billboards all around, if you were in a first-generation literate culture, what you would say is, "There's this terrible problem: As you go down the street, you're having your attention distracted by having to decode what this stuff is. There are all these symbols you have to decode. Meanwhile, you're not paying attention to anything that's going on in the street. Your attention is terribly divided." We know even neurologically that what actually happens is when you are deeply immersed in a literate culture, you end up with Stroop effects, where your decoding of print isn't attention-demanding in the same way. You're not doing it by serial attention anymore. In fact, you're doing it completely automatically and in parallel. It's something that we all worry about because we're in the position of the preliterate person. It's not at all obvious that this is somehow an intrinsic characteristic.

HILLIS: I'd like to bring this back to the AI part of the comment rather than the social part of the comment. If you look at where artificial intelligence is being deployed on a large scale, where people are spending a lot of money paying the power bills for doing the computation and things like that, they are mostly being done in the service of either corporations or nation states—mostly corporations, but nation states are rapidly catching up on that.

They are making those more powerful and more effective at working their emergent goals, and that is the way that this relates. So, when we think of these runaway AIs, we should think of them as not things off by themselves. They're the brains of these runaway things that are already hybrid AIs. So, they're the artificial brains or the artificial nervous systems of these things that are already hybrid AIs and already have emergent goals of their own.

LLOYD: This is why I disagree with you about this. Back in the 1960s, they would say, "Oh, kids these days, they're watching TV five hours a day. It's just horrible." Though I enjoy preparing for the grumpy old man stage of my life, and I like practicing that, I do think that if you look what these AIs are being devoted for, the primary use of them is to get people's attention to web pages.

HILLIS: Whether it's attention, or dollars, or votes, it almost doesn't matter.

JONES: The designers will tell you that they're using the lowest brainstem functions. That's part of the problem. They'll tell you they're racing to the bottom of the evolutionary channel as quickly as they can.

HILLIS: If there's anything valuable that is valuable to them, they will use this power to get it. There will be problems with that, and there will be limits on that and so on—you're pointing out some of the limits in getting attention—and there will be limits in their ability to get money, and their ability to get electric power and so on, but they will use all of these tools to get as much of it as they can.

GOPNIK: But again, Danny, my challenge would be, is that any different than it was for Josiah Wedgwood in 1780?

HILLIS: Yes. It's a tip in power.

GOPNIK: It seems to me you could argue there was much more of a tip in power if you're considering the difference between being around in 1730 and 1850.

HILLIS: For example, for the East India Company, they couldn't establish a policy and monitor that everybody did that policy. Google can. Google can do that.

GOPNIK: That's exactly what people at Wedgwood did. That was part of the whole point of investing industry, inventing factories was exactly doing that.

HILLIS: But in fact they couldn't do it very effectively.

JONES: East India had to translate itself to a language with an army, which was the British Empire. So, there are meshes between corporations and governments that we have to worry about, like the one we have right now.

GOPNIK: No. I'm not saying that we don't have to worry about that or there isn't power. The question is why is it that you think that *this* is a tipping point? It looks like there's this general phenomenon, which is that you develop these transindividual superintelligences, and they have certain kinds of properties, and they tend to have power and goals that are separate. All that's true but we have a lot of historical evidence, and it might be that what's happening is that there's more of that than there was before. But why do you think that this is a point at which this is going to be different?

HILLIS: There could be a tipping point. I'm not sure exactly now. What I am saying is that there's an explosion of their intelligence. These explosive technologies, which are driven by Moore's law and things like that, are being used to their advantage. There are very few examples where they're being used to an individual's advantage. There are lots of examples where they're being used to the advantage of these hybrid emergent intelligences.

LLOYD: That's a very good example, because between 1730 and 1850 the life expectancy and degree of nutrition and height of the average person in England declined because they were being taken out of the countryside and locked into factories for ninety hours a week.

GOPNIK: That's why thinking about these historical examples is helpful. If you think about the scaling difference between, say, pre-telegraph and train, so if you think about the difference in scale between the communication that you could have before you had the telegraph and afterwards and before you

had the train and afterwards, for all of human history the fastest communication you could have was the speed of a fast horse.

HILLIS: Yes. It made a big difference.

GOPNIK: Then suddenly you have communication at the speed of light. It seems to me there's nothing that I can see in what is happening at the moment that's different.

HILLIS: I realize what our difference is. I think of that as now. When I'm saying this is happening now, I'm including railroads and telegraph. This moment in history includes all of that, so that's the thing that's happening right now.

GOPNIK: That's essentially industrialization.

HILLIS: I'm not categorizing it. Industrialization focuses on the wrong aspect. A lot of things happened at once and you categorize them, but the particular thing that is interesting which happened at the same time as industrialization was the construction of an apparatus of communication of symbols and policies that was outside the capacity of a human mind to follow it. That's the interesting thing. There are many other aspects of industrialization, but that's the thing that's happening now, and computers and AI are just that going up on an exponential curve.

GALISON: Seeing this moment of increased poverty and stagnation of wages for a big sector of society, and enormous increase of wealth within a concentrated group, and the consolidation of industries like Amazon and others is something that does represent the sharp edge of that increase. It's not just a simple linear continuation of what went before.

In the post-World War II era, there was a sense that people were able in families to go to college for the first time, to get loans—at least if they were white—and that meant that you had a big class that had increased expectations and increased income. We're seeing the echoes of what happens when that stops when you're basically not bringing new people into the college system. You're not giving them increased stakes and homes and real estate and things that increase in value. We're at a tough moment.

GRIFFITHS: There's an interesting argument about something that's different, which is one argument that's often made by the technology companies is we're not doing anything different. This is something that's been done in the past, and we're just doing it better, but there is a case that you could make that doing it better is different. The objective function is the same, but you're doing a better job of optimizing it, and one consequence of that is that you get all of the unforeseen consequences of doing a good job of optimizing that objective, which may not have been clear when you were doing a bad job of optimizing that function.

In machine learning we talk about regularization. Regularization is forces that pull you back from overfitting on your objective, and you can think about not being able to do a great job of optimizing as a form of regularization, but it's helping us to avoid all of the negative consequences of really optimizing the objective functions that those companies have defined for themselves.

GALISON: They say we're doing the same thing, but they also say we like to break stuff, and breaking stuff often means breaking the income of working-class people.

GRIFFITHS: Yes, but it's enough that doing the same thing better is the thing that then reveals why it's bad to do that thing.

HILLIS: If you go back to the other perspective and say, "Is a single cell better off being a part of a multi-cellular organism that they can't perceive as living in a society that they can't perceive?" I would argue that it's a mixed bag, but generally they are.

GOPNIK: Right. That's right.

HILLIS: So, I'm optimistic in that sense.

GOPNIK: If you think of the train and telegraph is the inflection point, the individual achievement of goals didn't just get better but got exponentially better.

HILLIS: Again, I'm not seeing that as an inflection point. We're going through a transition. We're in the middle of a transition from going from one level of organization to another level of organization in that process. For instance, individual cells had to give up the ability to reproduce. They had to delegate it.

WILCZEK: That's a lot.

HILLIS: We will lose some things in that process. We'll gain some things in that process. But all I'm mostly arguing for is that we're spending too much time worrying about the hypothetical; it'd be better to look at the actual.

FREEMAN DYSON: The most important thing that's happening in this century is China getting rich. Everything else to me is secondary.

IAN MCEWAN: One aspect of humanizing let's call them robots, AI, whatever you like, would be to tax them as humans. Especially when they replace workers in factories or accountants or white-collar jobs and all the pattern recognition professions. Then we would all have a stake.

AXELROD: That's an example of where we may have passed the tipping point. The corporations are now politically powerful enough to keep their tax rates low and not only that, but the billionaires are powerful enough to keep their tax rates low. Inheritance tax, for example

MCEWAN: This is why we need to resist the point at which, perhaps in fifty years' time, vast sections of the population are only going to be working ten or fifteen hours a week, and we might have to learn from aristocracies of how to use leisure: how to hunt and how to fish, how to play the harpsichord. In other words, it's perfectly possible that anyone who speaks of retirement—and we were talking about this in a break—how busy you could be doing nothing. But somehow, we have to talk of distributing wealth and function here.

HILLIS: Bob's point is this is a sense in which the rubber meets the road where taxing corporations, that window has passed. We've lost that. They now have more power than individuals do in influencing the political system. So, there's an example of where the train has left the station. We're now in

a post-individual human world. We're now in a world that is controlled by these emergent goals of the corporations. I don't think there's any turning back the clock on that. We are now in that world.