Copyright © 2024 By Edge Foundation, Inc. All Rights Reserved.



To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves.

https://www.edge.org/conversation/david_chalmers-the-language-of-mind Printed On Tue November 5th 2024

CONVERSATION: SPECIAL EVENTS

The Language of Mind A TALK BY David Chalmers [8.8.19]



Edge Cast

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 codirector 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's** *Edge* **Bio Page**

Possible Minds Project

THE LANGUAGE OF MIND

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 mindbody 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.

What's Related People



David Chalmers
Professor, Philosophy & Neural Science; Co-Director,...

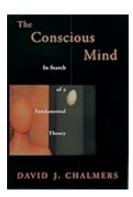
Video

The Language of Mind

Beyond Edge

David Chalmers' Website

Books



The Conscious
Mind: In Search
of a...
By David
Chalmers
Paperback

Tags
artificial intelligence
consciousness
mind
philosophy

Conversations at Edge

The Mind Bleeds Into the World

Mentioned

Daniel C. Dennett, Nick Bostrom, Andy Clark, Jaron Lanier

[1.24.17]

John Brockman, Editor and Publisher

Contact Info:editor@edge.org

Edge.org is a nonprofit private operating foundation under Section 501(c)(3) of the Internal Revenue Code.

Copyright © 2023 By Edge Foundation, Inc All Rights Reserved.