Brain and Mind

May 14, 2004

Christof Koch, PhD

Towards the Neuronal Basis of Consciousness

Introduction by David Cohen

David Cohen: Our third speaker this morning is Dr. Christof Koch, who holds the Lois and Victor Troendle professorship of cognitive and behavioral biology at the California Institute of Technology. He is also professor of computation in neural systems at Caltech. Born in the Midwest, Christof had a peripatetic youth, lived in Holland, Germany, Canada, Morocco. It was in Morocco that he graduated from the Lycée Descartes in—I'm sorry—he studied physics and philosophy at the University of Tübingen in Germany, where he earned his PhD in 1982, and after four years on the faculty at MIT, he moved to Caltech where he has remained.

Christof has published prolifically, both articles in books on the neuronal basis of visual perception, attention, and consciousness, and has had a long-standing interaction and collaboration with Francis Crick on the dialogue about consciousness. He's boldly engaged in an experimental program to study consciousness, the topic of his latest book, *The Quest for Consciousness: A Neurobiological Approach*. The book was reviewed in one of the issues of *Science* last month and I'll offer you the final sentence, "*The Quest for Consciousness* is a brave attempt to fuse the best of scientific thinking with one of the central aspects of human existence." The title of his talk this morning is "Towards the Neuronal Basis of Consciousness," and he will describe a two-pronged approach involving psychophysical and fMRI studies of humans, and animal studies that combine behavioral methods with single-cell recording and pharmacological and genetic interventions. It is my great pleasure to present to you Dr. Koch.

Christof Koch: Thank you very much. Actually, I cut half my talk yesterday night, so I'm just going to talk about some electrophysiology and some behavioral experiments in mice. Most of the time we are all conscious, hopefully you're still conscious, you haven't gone to sleep yet, and so what . . . I mean the state of being conscious, of being conscious of my voice, of being conscious of colors, or being conscious of pain or pleasure, of being angry or being you, those are all simple aspects of our existence. And by and large science, for practical and methodological reasons, has not considered those, even though they are so central to our—I mean they're natural phenomena, they do seem to occur, and we'd like to understand the scientific basis of these.

Now my talk's divided in three parts, so the first part is some conceptual work, some conceptual [inaudible] work that Francis Crick and I have done over the past twenty years, to construct a framework for how we'd like to think about the neural basis of consciousness. So as John Searle is going to tell you in the next talk, consciousness or the mind-body problem at large—it's one of the oldest problems in physics or metaphysics or philosophy, and of course the ancient Greeks have had a lot to say about it. At the heart of that problem is the problem of qualia, which is, how is it that a physical system like my brain or your brain undoubtedly, or the brain of a monkey or [the] brain of a fly undoubtedly are . . . how is it that some physical systems at certain times of their life have these subjective states? Not always. Not when I'm in deep sleep, not when I'm under anesthesia or not when I'm dead, presumably, and not all of—I mean not all complex systems have those states. My gut doesn't have them, although it's very complicated. I have an enteric nerve system down here in my gut, contains between 50 and 150 million neurons, and there's very little evidence that by itself is conscious, or many other systems both in nature and in our artifacts that we construct are presumably not conscious, so what is it about a subset of them that gives rise to these subjective feelings?

And this has been hotly debated among thinkers and philosophers and scientists in particular for the past twenty years. We think the most central aspect of that, what philosophers call *qualia,* which are the elements of consciousness—the red of red, and the painfulness of pain, philosophers refer to those as qualia—how do qualia rise out of the firing, out of electrical activity, synaptic activity, metabolic activity in the brain? For now, we prefer to leave that problem aside, for tactical reasons. We like to concentrate on the problem of the neuronal correlates of consciousness. By the correlates we mean, What are the minimal neuronal mechanisms in your head, or in your body in general? What are the minimal neuronal mechanisms that are jointly sufficient for any one specific percept, for any one specific conscious percept?

Now, Francis Crick and I focus, again for purely tactical reasons, for now we focus on sensory perception, on conscious sensory perception just because it's very easy to manipulate, in particular, vision. Visual psychologists have discovered over the last hundred years, and have perfected a whole range of techniques that allow us to systematically manipulate what you see and what's physically present. So effectively we can now do what magicians do all the time. Magicians in front of your eyes they distract you, they have a beautiful bikini-clad assistant next to them, and they distract you by using these well-honed techniques of attention distraction. They distract you and so they can make things disappear, although physically you're looking at it. And I'll show you an illusion like that, which we can also do, so we can now begin to manipulate the relationship between the physical stimulus that's on your retina in the case of vision, and the subjective percept in your head. And we can do that much better for vision than for other modalities, and we can do that much better for sensory modalities than we can do that, for example, for self-consciousness, which to many people is sort of central to consciousness. It's something very difficult to study. You can study it in humans but it's difficult. It's now being made possible with fMRI, but it's even more difficult to study self-consciousness in animals, which ultimately is where we have the best source of knowledge about the neural basis. So that's why we prefer to study visual consciousness. Although the belief is that consciousness is a feature of biological systems selected by natural selection over some, you know, tens if not fifty or a hundred million years. So it's likely that the central aspect of visual consciousness brings the system some advantage, some evolutionary advantage, that's also accorded to other aspects of consciousness, like consciousness for emotion or self-consciousness.

Neuronal Correlates of Consciousness

So once again the focus is on the experimental work, and there's a great deal of experimental work now that focuses on What are the minimal neuronal mechanisms that are sufficient for any one percept?

Our thinking sort of has evolved based on thinking of others, for example, physiologists like Bob Desimone and John Duncan, or going back earlier times to Hobbes, the idea is that your brain contains on the order of 20 to 50 billion neurons, and these neurons, they compete heavily against each other. They compete against each other, in other words, they can't all simultaneously be active for all sorts of reasons, partly because, of course, your brain would go into epileptic seizures, which obviously has to be avoided, and you have excitation inhibition among neurons, so you have coalitions of neurons that sort of . . . you can think of the brain a little bit like a Christmas tree, you have 20 billion or 50 billion little electrical bulbs on this and they all flash. These are the action potentials that Bill was just talking about in his talk. And so they all flash at different times, and they compete against each other, in other words, you have one group of neurons that suppresses other neurons. And ultimately consciousness arises out of the interaction among these groups of neurons. There's going to be one or several coalitions of neurons that correspond or that are jointly sufficient, that are sufficient for any one given percept.

One characteristic of conscious perception, as remarked upon already by William James at length, which is the fact that you typically can only be conscious of one or a few things at a time. The brain actively . . . there seem to be mechanisms in the brain that prevent two very similar things from being the focus of attention at any given point in time, you can only be conscious of one thing, and then of course you can rapidly switch your attention and your consciousness to something else. And so that ultimately, that's expressed by neurons that compete for each other. And so the trouble is our tools are very imperfect, we have tools, we have wonderful tools like Bill was telling you, electrophysiology, where we can listen to 1 or 2 or 50, or now with advanced technology, micromachining, we can listen maybe to 50 or 100 neurons, but we are sampling from a hundred neurons out of a sea of 20 or 50 billion, and so that's a practical problem that we're facing.

And so the belief—you have this coalition of neurons, and ultimately the winning coalition for any given point in time. And the winner of that coalition is . . . the representational content of the winner of that coalition of neurons is what you are conscious of. So it can be a voice and then it rapidly switches to an image and then it rapidly switches to the fact that you know your leg is itching.

Zombie Systems

As we know from our own personal experience, and this is, of course, a point that was much remarked upon by Sigmund Freud: Much of what goes on in our head bypasses consciousness; much of what goes on in our life we do totally automatically. And so there's this whole set of systems that Crick and I call Zombie Systems. These are highly attuned sensory motor systems that by the time you learn them effortlessly you do them automatically without having to think about them. You do them mindlessly in the sense that you don't have to think about them. So this includes, you know, in the morning you get up, you tie your shoes, you drive to work, and you type on the computer. Anything, sort of things like driving, like playing basketball, like climbing, like dancing, all these activities like moving your eyes, like moving your limbs, like reaching out and grabbing something, all of the things we know from experimental psychology and from clinical studies, are automatic. We do them at a very high level of proficiency; in fact, that's the point of training, that's why you train and train and train, so you don't have to think about them. And, in fact, very often if you do these activities so well, typically if you think about them, if you stop and think about [them] consciously it'll interfere with your performance. And evidence seems to be that the sensory motor system that is highly trained up, like including eye movements that typically don't have access to working memory, that if you use a system that requires working memory you then invoke a second set of systems, a system that seems to correlate with consciousness. So the claim is that you have this architecture where you have these two systems, on the one hand you have all these automatic sensory motor systems that control most of your life, and then you have this additional system that's much more powerful that allows you to do any arbitrary complicated task. And this is the system that empirically seems to be associated with consciousness. This system also has access to planning; in fact, that seems to be one of its key characteristics. That if you want to do planning, if you want to think, you know, suddenly there's a fire here, how do I get out? You know, how do I leave this building? I then have to bring to memory, I have to recall where's the entrance, how I can get to that entrance, that's a system that involves consciousness. And so the function of this system, Crick and I believe, is to plan. That's one of the key functions of consciousness, and that's probably one of the key reasons why it arose during evolution, to enable the system, to enable the animal, to do planning, to do things beyond the stereotypical response that the animal had learned.

This hypothesis has some anatomical correlates, in particular since we are arguing that consciousness is principally involved in planning, therefore we surmise that the neurons that underlie the consciousness, or the NCC, the neural correlates of consciousness, that they have to have direct access to the planning stages of the brain, which by and large are in the frontal part of the brain. If you look at the neural anatomy based on a monkey, we know scandalously little about the detailed neural anatomy in humans. Based on the neural anatomy in the monkey, if you look at one of the best explored areas in cortex bar none, which is the visual cortex that Nancy showed you already, it's at the back of your head. In fact, you can feel there's a little bump here at the back of your head, and your visual cortex is a little bit above it. We know that because if you get hit on the head there you may see sparks or flashes of light. That's what happened to cartoon characters in any case when they get hit there. So that's your visual cortex. This is your primary visual cortex, the first entry point of the visual output from the eye through the thalamus into your cortex. This particular area is also called *V1.* This area does not have any neurons that project into the forward part of the brain; in particular, it doesn't have any neurons that directly project into the planning stages of the brain. Therefore, we surmised ten years ago that neurons in V1 are not sufficient for visual consciousness, that they may be important for seeing, just like my retina is important. Clearly if I don't have my eyes anymore I cannot do normal seeing. I can still do imagery, I can still dream visually, but I can't see. So, likewise, neurons in the primary visual cortex are important in this sense, but they're not sufficient for consciousness. And consciousness has to arise from discrete coalitions of neurons in a higher part of the brain, in a higher part of cortex.

And this has interesting consequences. You can test this. And so our claim was that the NCC . . . that visual consciousness does not arise, is not . . . that this part of the brain is not sufficient to give rise to visual consciousness, that visual consciousness has to be generated in higher parts of the brain. And there's lots of evidence for that, in particular, the most striking is evidence from monkey electrophysiology, from recordings done by Nikos Logothetis and his colleagues, where you can see that the monkey responds. We do experiments where you manipulate, just as I mentioned to you, where you manipulate the relationship between what's on the retina of the monkey and what the monkey sees. The monkey doesn't see a stimulus, although physically it's still present on his retina, and you see literally millions of neurons are firing to this unconscious stimulus, it's a stimulus that the animal doesn't perceive and doesn't respond to because it's perceptually suppressed, yet there are millions of neurons in primary visual cortex that still respond to this unperceived stimulus.

And there's also some nice evidence from human . . . there's also now some recent evidence that other primary areas, like the primary auditory cortex or the primary somatosensory cortex—also you can have a lot of activity in those regions without being at all conscious about them. So it may be possible that none of the primary sensory cortical regions actually are sufficient for consciousness. You could say, Well if that's true, you know, that's sort of a minor detail—who cares?

Well that's, I think, very interesting because it suggests that not any neural activity is sufficient for consciousness, in fact, not even any neural activity in [the] cortex is sufficient for consciousness, that consciousness is something discrete, you know. There are two sets of ideas as John will tell you more about later. One sort of holds that it's impossible to assign the genesis of something like visual consciousness to any one specific set of neurons, that consciousness is a global, holistic, collective, gestalt-like property of the brain, and it's silly to think that it arises from discrete groupings of cells.

Now our intuition is based on sort of the model of twentieth-century biology, which is all about specificity. And you saw that yesterday, for example, in the talk by Richard Axel where you have amazing amounts of molecular specificity at the level of the antennal lobe in a fly. The same overall story is going to be true for cortex and for the generation of consciousness. There are going to be discrete biological mechanisms in the brain. They will involve discrete groups of subtypes of cells that talk to each other in particular ways, and they will give rise to specific types of conscious sensation. While other neural activity does not give rise to conscious sensation, and this is the activity involved with all these Zombie Systems, all these things when you're driving home, you're lost in thought, suddenly you realize you . . . sort of . . . you wake up and you're in your garage while you were thinking about your latest paper that just got rejected. And so here you had to do very complicated activity, we know that there's complicated activity, you had to stop, you know, you had to look at traffic lights, etcetera. We know this involves cortex, so again we know that it can't just be any cortical activity, it has to be specific types of cortical activity, you know, in a particular mode maybe, in a particular region of the brain that gives rise to conscious sensation.

Explicit Representation

The other thing that we think is absolutely essential for the NCCs, an explicit representation underlying every discrete conscious percept, like conscious percept of my voice or of these colors, it's got to be an explicit representation at the neuron level. But what I mean by explicit—this is a picture I've borrowed with Bill's permission from one of his earlier papers—so this shows you an explicit representation for depth and for motion in an area called MT or V5. This is a part of a cortex of a monkey where you have a whole—this is part of cortical layer here, so this is roughly two millimeters, and here you have an entire set of neurons from layer II to layer VI that all sort of more or less code for motion in this direction. And over here neurons code for motion in this direction or this direction or this direction. And they encode this direction of motion explicitly, in other words it's very easy for you as a postsynaptic neuron, as a neuron network or receptor, it's very easy for you to read out that information explicitly. Likewise, if you go in a higher part of the brain called inferotemporal cortex, where we know from monkey recordings a lot of neurons that encode faces, that encode faces in a very explicit manner. In other words, for a postsynaptic observer, a postsynaptic cell, it's very easy to decide based just looking on the output of these neurons whether or not a face is present.

You take the counterpart to explicit, it's an implicit representation. So everything I know about the visual world is already present in the retina, because that's the only source of information. In fact, it's already present at the level of photoreceptor. But things are not made explicit. At the level of the voltage in the hundred million photoreceptors in my eye, this information is not made explicit. Information for faces is only made explicit at a high-level stage. And the idea is that the key apparatus . . . is that the neural correlates of consciousness has to be based on such an explicit representation. Why? Because we're directly conscious of something, we don't have to—one of the remarkable features of consciousness is that I'm directly conscious of it. That's what it means to be conscious of something. And so therefore there has to be a direct neuronal counterpart, there has to be an isomorphism, in other words, between anything in my conscious representation, between any attribute of qualia and the underlying neuronal representation.

Then there's the ideas that our . . . we have to ask, How do we experience the world as evolving in time? Are experiences evolving continuously or do we experience this evolving discretely? Certainly if we introspect we experience a smoothly changing world. There is evidence that that's an illusion, that similarly, like the way you experience motion in a movie, which we know actually is incorrect, there's nothing that moves actually in a movie, right? In a movie what you have, you have a discrete image here, and then at 72-Hertz frame rate you have a discrete image here, here, and here, and if you do it quick enough that creates the illusion of motion. Likewise, it is possible that perception in the brain also occurs in these discrete snapshots. These snapshots can be a variable period between, let's say, 50 milliseconds and 120, 150 milliseconds, depending on all sorts of circumstances, the saliency of the stimulus, etcetera, that perception actually occurs in these discrete episodes.

We've written about this last year, and then we were struck by something that Oliver Sachs communicated to us, and he has subsequently written about in the *New Yorker*, what he calls cinematographic vision, which is under conditions of migraines. And he himself has experienced this, I think I have a slide of that—no. So this is sort of the metaphor, that the way you experience things is actually not continuous, but it's discrete. But if it's quick enough, just like in a movie, you can't tell the difference unless you have an explicit mechanism in your head that tells you, and that actively, explicitly signals this difference. So what happens in cinematographic vision? These are people who have visual migraines, and what they experience is—and here's this very vivid description that he himself has had—that the world is fragmented in time, that you see things. Like, for example, he describes when Oliver Sachs had this migraine, that he sees a nurse approaching him in the hospital, but sort of she—it's like the movie's run too slow, and that's in fact the description that most of the patients use when they have these type of cinematographic visual migraines, it's like seeing a movie run very, very slowly.

Several speakers yesterday and today mentioned attention. What is the relationship between attention and consciousness? Now different from William

James, we think that attention and consciousness are actually distinct states, that attention is a general set of mechanisms, there's both bottom-up attention that's inherent in the input, and there's top-down attention when I can sort of direct my attention in a trained situation—like in Bill's monkey—I can direct my attention to one or to the other stimulus, even though they are of equal visual salience. Attention is a set of selection mechanisms that enables me to take all the stimuli that compete for my consciousness, which is much more than I can process in real time, and to select a very small subset of those stimuli. And a subset of those stimuli, those are the ones that I'm then conscious of. And we know from a hundred years of visual psychology that there are many, many things outside there—we also know this from our personal experience—that I'm not conscious of at any given point in time. So attention is a set of neural mechanisms that selects, out of all this competing stimuli, you know, that are on my retina, that are on my body, that are in my cochlea, that are sort of in my internal imagery and representations, selects a subset of those and then a subset of those again are the ones that are consciously accessible. If they're consciously accessible then you can talk about them, you can access working memory. And that selective attention is necessary to what Richard Axel yesterday referred to as the binding problem. That if I want to recognize certain objects, particular objects I haven't seen before or if I want to combine property attributes of different objects at the same time, that's when I need selective attention, that's one of its key functions.

Styles of Neuronal Firing

Much has been made, including by ourselves, about different types of neuronal firing, so you may know electrophysiologists have characterized under different conditions if you listen to the way neurons talk to each other they seem to have different modes. And sometimes you hear neurons that just fire randomly, rat-tat-tat-tat-tat, rat-tat-tat-tat-tat, like a Poisson; in fact, a Poisson statistic is a reasonably good approximation of that. But then under other conditions you can also hear that neurons sort of fire rhythmically, in various frequency bins, particularly, there's this one called *40-Hertz firing* where neurons seems to fire with some sort of periodicity in the 20- to 30-millisecond range, roughly in the 40-Hertz range. And then furthermore what you can see under certain conditions, neurons from two separate neurons don't just fire independently of each other but seem to appear to fire in synchrony. This is called synchronous firing. And many people have argued that one of the key properties underlying any neuronal representation of consciousness is the fact that neurons that code for conscious—or that express or that are jointly sufficient for conscious—percept are the ones that fire in synchrony. The evidence for that based on monkeys is rather inconclusive. There's some evidence, the evidence is not very good. The best evidence indeed for the importance of synchronized firing comes from the olfactory system in insects, the evidence of Gilles Laurent, so it is more likely that synchronization and synchronized firing is important for certain types of conscious stimuli, in particular when they are competing into each other. In fact, there's now some evidence from Bob Desimone that 40-Hertz synchronized oscillation may be important for biasing correlation. So you have two stimuli—both compete for attention—you can only be conscious of one of them, so you somehow bias and you tend to one stimulus, not the other, and that one of the neuronal signatures of that bias is actually synchronized oscillatory firing. But that this type of firing therefore often occurs with consciousness, but that it may not be absolutely necessary. In particular when the conditions like you often have in a lab where you just put a single stimulus out on a monitor and the subject is looking at a single stimulus when there isn't a lot of . . . when you don't really need too much attentional-selection bias because there is really no other competing stimulus. So in other words, under certain conditions, there may be this relationship between synchronous firing and consciousness, but it probably does not hold in general.

Types of Behavioral Assays

So what is the experimental program we advocate? Well first of all it has to be emphasized again and again that we think that consciousness is an empirical problem, maybe not a purely empirical problem, but consciousness ultimately is an empirical problem that's amenable to sort of the normal scientific method that we've used so successfully over the past few hundred years. What you need is a series of assays in humans, and particularly in animals, you need a series of behavioral assays where you can be sure that the subject is conscious. Now for us that's sort of trivial, I mean I'm certainly conscious because I can feel that, and assume that most of you, if not all of you, are conscious. I just mean those of you who have fallen asleep, not that you're zombies. But in an animal we . . . I mean most neuroscientists would assume because of the great behavioral similarity between sort of a typical subject, if you take a typical undergraduate subject and experiment and you compare that against training an animal in a similar task then you'll find a great deal of—particularly if these tasks do not involve high-level, you know, knowledge, obviously animals don't talk, but involve simple things like vision—you find a great deal of continuity between the behavior of animals—in particular nonhuman primates like macaque monkeys—and humans, the brains is very similar. If I give you a little cubic millimeter of a macaque brain, a little cubic millimeter of a human brain it's very, very difficult to tell them apart. We've got much more brain, but the basic structure, the basic neural cell types, etcetera, are very similar. So therefore, by and large, neurobiologists sort of assume that animals share at least some aspects of consciousness with us, probably not self-consciousness. Self-consciousness seems to be something that's probably highly elaborate in us that maybe only chimps or some, you know, the great apes share with us, but certainly the way most animals, certainly I would believe that most mammals see the world, hear the world, smell the world is probably very similar to us, up to and including the conscious perception of these stimuli. So what you would like, you'd like to have a set of behavioral assays that you can use in monkeys or even other animals that are more amenable to genetic interventionist protocols, such as a mouse, or maybe even if you want to go lower, maybe even something like a fly. And that ultimately – and you need that because doing fMRI and doing single cell, doing psychophysics in humans is great, but ultimately you need to intervene with the system, you need to perturb the system selectively. Cortex is by far the most complicated system for its size in the known universe, and the only way we're going to unravel the circuit is by sort of taking, you know, going inside the brain, taking the circuit apart and then putting it together again to understand the genesis, origin, and function of consciousness. And obviously we can do that in animals.

One set of assays seem to involve a whole set of different—in humans and in patients—tests that involve keeping information online for at least a few seconds. There's really no convincing evidence that you can do that sort of task, you know, I give you numbers or, you know, a patient has to remember the orientation of a stimulus or the color of a stimulus, there's no evidence that you can do that, or patients can do that, you know, if you have to keep this information online for a few seconds without having access to consciousness. So probably a good sort of set of, you know, like a [inaudible] test, a good set of behavioral assays would involve tests that involve the subject having to keep information online for at least a few seconds.

So we are pursuing a number of paradigms, mainly psychophysics and fMRI, but yesterday night I decided just to focus on two, given the brevity of time. So one involves mice and the other one involves humans.

Recording Single Neurons in Conscious Humans

So almost everything we know about electrophysiology we know from electrophysiology of animals, for obvious reasons. Now there are, occasionally, rare occasions when you can record the response of individual neurons in humans—that's during surgery. So we work with a neuroscientist and a neurosurgeon, Itzhak Fried at UCLA, and the background is that these are epileptic patients, these are patients that are resistant to conventional pharmacological treatment, the drugs don't work anymore on them. And then what neurosurgeons do, it's done throughout the world, it's a quite successful operation, then you go in and surgically remove the part of the brain that gives rise to the foci, which more often than not tends to be in the medial temporal lobe. You remove, for example, a piece of hippocampus. And then, you know, the incidence of seizures is dramatically lowered. In some patients the incidence goes to zero, in most patients it's dramatically reduced. Now in a subset of these patients, roughly thirty patients, on the order of one patient every month, from the outside looking at the etiology, looking at the behavior, looking at the EEG, or even looking at the structural MRI is not sufficient to tell you where is, actually, the foci in the brain. So then what you do, you put in electrodes, you put in up to 12, ten to 12 of these microelectrodes, so these are big electrodes, these are 1.2-millimeters thick, these are polyurethane, very flexible polyurethane, and then you have these platinum-iridium leads. So here you can see one inserted into the hippocampus of this patient. And then the cortex is sealed up, the burr hole is sealed up, and then the patient has, like I say, eight to 12 of these electrodes in his or her head, and then is monitored on the ward for let's say three to five to seven days, it's monitored 24/7. And then when the patient has seizures, in fact the patient's released after he has three to five seizures, that's typically the number that's sufficient to enable the neurosurgeon to then triangulate where the foci is. The electrodes get taken out, and then that foci is taken out and that's the end of the operation.

Now, in fact what Itzhak Fried and his collaborators did already a number of years ago, they hollowed out the inside of this polyurethane and now they inserted these metal wires. These are just like Bill showed you; now, conventional microelectrode isolated until the last forty micrometers, and they're very thin platinum-iridium metal wires essentially. So typically we use nine wires here and there are eight to ten of these, so we have on the order, and we have 64 preamplifiers, so we can now record from on the order of thirty to fifty neurons for days at times, for two to four days, in the head of a patient. So there are some advantages and some disadvantages to this scenario: The advantage is with the permission of the patient we can do experiments. We can directly query the person about his or her conscious state, we can directly ask you did you see that, did you not see that? And also the patients can do things without training them for many, many months. In a typical monkey experiment you have to train the animal for a very long time and reward him suitably in order to get the monkey to perform or the animal to perform the task that you want the animal to perform. In humans that's of course much easier. The drawback is it's not nearly as well-controlled as in an animal situation, and there are many things we can't do. We can't, for example, get the person to do the same experiment hundreds of times, just because the patient isn't willing to do that and will literally go asleep.

[A portion of the transcript including unpublished results has been removed at the request of the speaker.]

So then what we can do if we record from any of these neurons, like I said, typically now we get between thirty and fifty of these neurons, although many of those are not visually responsive, we can do what people called decoding. And in this case it's just a very superficial linear discriminant decoding, so, in other words, we can ask I'm now listening to, let's say, seven neurons from this one batch of electrodes, all of which seem to respond to various visual images. So on this one trial where I showed an image for one second, how well can I—using some objective criteria—how well can I decide which of the twenty stimulus was present? So once again I show the patient, let's say, twenty different pictures, and now I want to do decoding, and the decoding question is based on the firing of just this set of neurons during this one trial. How well can I discriminate whether it was the picture of, you know, the picture of a person or the picture of that animal, or whether it was the picture of the spider, or picture of the eagle? So we can get these probabilities. This is in different sessions in these patients. This is from the last patients where we plot some of the data. Let's see, the blue is chance level, so if I'm just guessing this . . . so here's a probability, these are just different trials in different patients, this is a probability from 0, 20, 30, 40, 50, 60, 70 percent. One hundred percent is perfect. Chance here varies because it depends on how many stimuli we show. So for example in this particular case chance is 50 percent, so if I'm just guessing I would get 50 percent. And so this is what a . . . the light blue is the performance of a neuron if I look at all the spikes over the one second when the image was present and the one second following the image. So I can do much, much, much better than chance. I can do even much better if I just focus on the burst of spikes. Very often we have cells that when they respond, particularly when the person recognizes the object, they seem to fire in bursts, and in a burst of spikes, a quick, you know, five to seven spikes in one hundred or two hundred milliseconds, that burst of spikes always occurs between three hundred and six hundred milliseconds. So we just look at that information, we just look at the action potential between three hundred and six hundred, then we can do very often much, much better than looking at all the information. In other words, the information is not just uniformly spread across the interval, but it seems to be specifically, there seems to be much more information in this one particular burst interval.

[A portion of the transcript including unpublished results has been removed at the request of the speaker.]

Here's a cell that seems to be selective to very different pictures of Clinton, doesn't respond . . . let's see here, here, and here, so again this is the response to stimulus where it came on here, came off here, and the horizontal bar tells you the background rate. So the neuron fires on average maybe two to three times a second, but clearly fires very, very strongly to these three different pictures. It's a remarkable invariance property, because if you look at the level of pixels, you know, the relationship between here, which is a gray-scale image where his face takes up almost the entire image, to here, where it's color and is much smaller, to here, where the face is even smaller, and you know, there are other individuals present, is really . . . so at the pixel level these images are very different, but the neuron seems to respond in a very invariant manner to it. We find that quite a bit in these areas. Typically we have out electrodes in the medial temporal lobe, which is amygdala, hippocampus, parahippocampal gyrus, and interrhinal cortex.

Motion-Induced Blindness

So this is one particular experiment that we did: What you should do, you should look at the fixation, you should fixate and not try to move your eyes, and just look at the fixation cross in the center. It's a nice illusion discovered by Bonneh four years ago. It works best if you don't move your eyes, if you keep your eyes as still as you can on the cross. There's no way we can make this darker, right? Do some of you see something—yes, no? Oh thank you, excellent. You see things disappearing? Okay, so this is one of these illusions I mentioned, this is called motion-induced blindness, like I was saying, was discovered four years ago by Bonneh et al, and it's one of the many illusions that visual psychologists have now to manipulate what you see, to manipulate the relationship between what's physically present on the screen and what you see inside the privacy of your head.

Here the point is, although those yellow squares are present all the time—they're all the time present on the screen—sometimes you see them, sometimes you don't see them. When you're seeing you're conscious of them, they're yellow, they remind you, I don't know, of the yellow sun, they remind you of yellow sunflowers of Van Gogh, I don't know, whatever else you have, what memories you have associated with yellow. You're conscious of them, you can talk of them to your neighbor, you can use them to do planning. And when they're gone, they're just gone, you don't see them anymore, you're not conscious anymore. So the question is where's the difference in your head? It's a very simple question, right? Where is the difference in your head? What's the difference in your visual cortex between when you see them and when you don't see them?

And, for example, you can use these sort of stimuli, this is just one of many such illusions, like rivalry flash suppression. All are instances of that, where you can take the footprints of consciousness, right, because now you can ask every time when you consciously saw the yellow spots, you could do it in fMRI, and has been done in fMRI, Nancy mentioned that, she has done something like this, when you physically see the yellow is there a region of the brain that's active in fMRI, or are there single neurons that fire when you see the yellow? And if you don't see the yellow but the yellow's still present on your eye, are the neurons, is that neuron, does the neuron stop firing? If it does then it has its close correlation between visual consciousness and firing. Of course that's just a correlation, it's not a causation yet.

For example, you can show, we've shown recently, that you can now—I mean you all know what a—I assume most of you know what an afterimage is, right? So, for example, if you stare at the yellow for a long time and then I flash you a black screen you're gonna see a ghostly blue image superimposed. Why blue? Well because yellow gives rise to blue afterimage. If you look at a red image for a while you get a ghostly green afterimage. For example the extent and the duration of your afterimage doesn't at all depend on whether you consciously saw the stimulus. So in other words you do not need to see a stimulus in order to get an after effect. Certainly you don't need to see a stimulus in order to get an afterimage of the same duration and the same intensity, and you also don't see a stimulus to get an orientation-dependent after effect. All those things preceded consciousness. So what that tells us there are discrete stages in the visual system, in the visual hierarchy, and that at one stage you have the neuronal correlate of these after effects, for example, the afterimage that's probably in the retina, or orientation-dependent after effect in V1, and that visual consciousness has to arise at a higher stage. So again that's important because it tells you it's not just one holistic thing, but that there are discrete processing stages and consciousness seems to arise at or beyond a particular processing stage in the brain.

Flash Suppression

So now, in principle, we can do the same in a patient. So this is a similar version, it's just more difficult to demonstrate in an audience, which is why I'm using the other one. Flash suppression is a similar illusion, was discovered and characterized by Jeremy Wolfe at MIT, and essentially involves the following: I show you an image in one eye, so let's say you have an image of my hand in one eye, and then I don't know, and you clearly see that, and then I flash up in your other eye I flash up the image of the watch. Now both the image of my hand, if I do it carefully the image of my hand and the image of my watch are simultaneously present, but perceptually the new input trumps the old input, and what you see is only the watch, you don't see the image of the hand, although physically the image of the hand is still present in my right eye. Okay?

And there were some beautiful experiments done by Nikos Logothetis and David Leopold in a monkey where they showed already in a monkey that there are cells early on, in primary visual cortex, that don't care about—that seem to fire to the stimulus whether or not the animal saw the stimulus, whether or not the animal behaved to the stimulus. So you can get a monkey to train in flash impression just like you can do, you know, its behavior is very similar statistically to the behavior of a human, and you can see in a monkey that cells in V1 fire—fire very vigorously, even though the physical stimulus is present. So again, that supports the idea that primary visual cortex, that's not where visual consciousness—the neurons in primary visual cortex are not sufficient for visual consciousness.

So we can do the same in patients, so this is one neuron. So here we show Curly, we showed Curly because we found neurons that responded to Curly in this gentleman—well, in fact, this is the neuron that responds to Curly, so here we put Curly on, nothing in the left eye, the patient recognized Curly, incidentally, and here you can see—again you see this burst of activity around three hundred milliseconds, and the person sees Curly. Now you flash on a grating in the left eye. Perceptually this grating suppresses the image of Curly and what you see—you don't see a superimposition, you actually only see, it's quite striking, you only see grating. It's very reliable, easy to set up in the lab. And after a suitable delay, this response here [is] statistically no more different, statistically not different from the response here. So in other words, the neuron behaved statistically, at least to the best of our statistical abilities. The firing here is no different than the firing here.

Here we do the opposite case, we show the grating, the neuron doesn't fire to the grating. This neuron really tends to like only Curly, this is the only stimulus we found that this one cell happens to like. The person sees Curly, you flash up Curly in his right eye, perceptually this image suppresses that and you see Curly. Now notice here, this input and this input are the same physical input, the only difference is here the person says—I mean I know it, we asked him—sees a grating, and here, he sees Curly. So again, here the neuron fires very strongly. So in this case the neuron correlates with the visual consciousness of the person. So we can do that in category-specific cells, we can do that in cells—here we averaged a lot of cells in a couple of patients that are category selective, here we did it for neurons that are selective only to individual images, so I can just show you the average response. So when I put on the preferred stimulus of that neuron, like Curly or Clinton, the person saw the stimulus and the neuron fired, then here the preferred stimulus was still present, the stimulus that the cell responds to it was still present but was perceptually suppressed, and the neuron fired only weakly. And here the opposite is the case, that here the stimulus isn't present, the neuron doesn't fire. Here the stimulus is present and is perceived, and the neuron fires very strongly. So perceptually this condition here, and this condition are the same. Here you only have the preferred stimulus present, here you have the preferred stimulus present in one eye that suppresses the other stimulus in the other eye, and you can ask, Is anything different about the neuronal response between here and here? And we did that and the answer is *no,* neither the duration, nor the amplitude in terms of the number of spikes, nor the peak response is different between those two cases. So again, statistically, we cannot tell apart based on the firing of this neuron; sorry, the neuron cannot tell apart, or at least the spiking of the neuron does not distinguish between the two situations that are phenomologically the same. That I have only Curly present, or I have Curly and the other stimulus present, but Curly suppresses the other stimulus. In both cases, to the observer, they look the same. I see Curly, and the neuron signals that likewise.

So now this is, of course, just correlation, like a lot of electrophysiology. All I can tell you is that there's a nice correlation between what the human said he saw and the behavior of the neuron. By the way, this was true for two-thirds of the neuron in this part of the brain followed the percept, one-third of the neuron just didn't fire at all or fired much reduced when two stimuli were present. We never saw a scenario where a cell fired to a perceptually-suppressed stimulus. In other words the unconscious, you know, the Freudian unconscious if you want, wherever it is, it's not present in the firing rate of neurons in the medial temporal lobe.

From Correlation to Causation

So there's this nice correlation. Ultimately, you want to move to causation. Now that's not impossible in human neurosurgery, and possible scenarios one can imagine among other stimulation, occasionally—or, in fact, quite commonly—neurosurgeons do stimulate parts of the brain during surgery to make sure, to understand where they are and where the language areas are and where the motor regions are. So it's not implausible that you can think about a protocol, there are a number of obvious ethical and practical questions involved, where you can think of a protocol where you can directly stimulate, you know, a bunch of neurons that seem to code for faces or for animals to try to see: Can you actually switch—in a reliable way, can you induce a percept or can you reliably switch the percept of a human in order to begin to take the jump from correlation to causation?

Of course in animals, this is much easier, and so that's where we're exploring a totally different paradigm. So this is work we've been doing with my good colleague David Anderson at Caltech, and Michael Fanselow at UCLA. And David and I have two postdocs, C. J. Han and Colin O'Tuathaigh, who actually did all the work. So this is based on a paradigm by Clark and Squire. So there are many forms of associative conditioning, some—have been, since a long time—require conscious awareness of the relationship between the CS and the US, other ones do not. In this *Science* paper that they wrote five years ago, they looked at—in humans—they looked at eye-blink conditioning, and they argued that this form of conditioning where you have a tone, beep, and at the end of the tone you get a puff of air to your eyes. Now that's very annoying and you blink. If you do this a hundred times and you just get the tone you immediately start blinking, your deep cerebellar nucleus, your brain has learned to anticipate, when this tone comes I've been conditioned to expect a puff of air to my eye and I blink. This is a more complicated form of conditioning, it's called trace conditioning. Now you have a tone, *beep,* you can also make it long, they've also tested that, you have a tone and then you have an intervening trace period of one second, and only then does the puff of air come. Now this is, as I mentioned to you before, you know, as an assay to involve consciousness, this is one of them, I think, because it involves this intervening period. And so now the animal, the human, there isn't just the tone and the puff of air but there is an intervening period and you have to keep it dynamically online. And they showed some nice evidence in humans that this requires awareness, this requires, here people need to know that there were tones and there were puffs, and the tones always preceded the puffs in order to express conditioning. Here, whether or not people were aware of this made no difference; here, the subjects were always conditioned. We repeated the same thing in humans using shocks, we did it using electroshocks because we'd like to move to mice, or we're moving to mice, I'll show you that now, and in mice you can do field conditioning, this is called field conditioning, much more reliable, it takes many fewer shocks than puffs of air.

So what we did is the following: So this is now in mice. So we have delay conditioning, here we have a tone, beep, for 16 seconds, I'll show you a movie, and then the floor of the cage is electrified. That's delay conditioning. And here's trace conditioning, here there's an eighteen second trace period between the end of the tone and the shock. And so the animal has to keep this dynamically online. Then we try to distract mice. One of the ways Larry Squire showed that you require attention and awareness in humans—he distracted humans. So we do the same and we try different things, and here we distract them by flashes of light, a bit similar to what Eric Kandel was telling you yesterday about in his case. So here there are flashes of light, and then we do the same thing, we do trace, delay conditioning and trace conditioning. So these mice have never been shocked. This is the very first time we shock them.

Okay, so it's working now. So there's this tone, which you didn't hear, because I didn't put on the audio, and you'll see what happens. There's these flashes of light, and soon you'll see what happens to these critters. Okay, so now for two seconds the floor was briefly electrified. And now, you know, they're all very nervous. And we do this six times. And this was the very first time. We do this in day one. Then on day two we take them into different context, there's no context-dependent conditioning, and then we test them. So here you have two sets of mice from one conditioning and two sets of mice from a different paradigm. One set was distracted, the other one was not distracted, you'll see which one. And the measure of conditioning we use is freezing, how much do the mice freeze, and you'll see—okay, there's this tone—okay, there's no tone here. Okay, so I can just tell by the light, the tone is on and you can see these mice totally stopped moving, the mice are somewhat reduced moving but they still move, so this is called freezing. And you can measure the amount of freezing by doing behavior using a videotape and measuring every second this mouse is frozen, that mouse just stopped freezing, here the mouse stopped freezing but then it goes back to freezing, these mice freeze much less. The difference is these mice on previous days were exposed to the trace-delay paradigm with distracter, these were exposed to the same shocks, the same tone, but here they were distracted. So here you can see the behavior average, so this is delay conditioning with and without distracters, statistically there's no difference. Here's a very significant difference between the mice that trace conditioned that were . . . so these were not distracted and these were distracted.

So the other way, as I said, we now want to begin to move to interventionist protocols, so what we then did—and I guess I'm running out of time—we did pharmacological lesion to remove the anterior cingulate, which we know from our functional imaging experiment is a part of the brain in humans that's involved specifically in trace conditioning, but doesn't seem to be so much involved in delay conditioning. And if you do that, if you remove the ACC in these animals, if you remove this in animals then you get this very nice behavior. So these are the normal animals or with sham surgery or V1 surgery, these are the animals with ACC lesion. Makes no difference whatsoever in delay conditioning, but essentially eliminates all of trace conditioning as compared to shock only. And it doesn't interfere with context conditioning.

Conclusions and a Warning

So, to finish, what we could show here in the mice—we have a nice model, similar to what Eric Kandel has, we have a nice model for attention, possibly if you believe the link to humans with awareness—in a sense that we can show two forms of conditioning: trace and delay conditioning. If we distract the animal it seems to interfere specifically with trace but not with delay conditioning. And if we remove a part of the brain, ACC, that in humans in involved in something similar, you can again eliminate trace conditioning without interfering with delay conditioning or context-dependent conditioning.

Let me give you one last slide. So people say well, this is all very fine, nice and fine and you know, this program of finding the neural correlate of consciousness is interesting, but you know, will that explain it? So let's say it's layer V cells in inferotemporal cortex that project to prefrontal cortex and back that are sufficient for consciousness. How does that explain something? And there's this wonderful quote I found, and this is Bateson, who was a very famous English geneticist, and he reviewed a book of Thomas Hunt Morgan, who was at the time, here, I believe, at Columbia, during the war, and in this book he wrote about his evidence based on flies, that genetic information was stored along one-dimensional strings. And so this is what he wrote, "The properties of living things are in some way attached to a material basis, perhaps in some special degree to nuclear chromatin." We know that's true now. "Yet it is inconceivable that particles of chromatin or of any other substance, however complex, can possess those powers which must be assigned to our factors or gens [sic]." That's his spelling. "The supposition that particles of chromatin, indistinguishable," that's of course incorrect, "from each other and indeed almost homogeneous under any known test can by their material nature confer all the properties of life surpasses the range of even the most convinced materialism." The trouble here was that they thought they understood chemistry. And so by their test at the time in the early 1900s they couldn't distinguish different strings of one-dimensional information, in fact they couldn't even imagine, they didn't even have the concept of sort of specific-marker molecules. Hemoglobin wasn't characterized until later. And so they couldn't imagine the amazing complexity, the amazing amount of information you can specify in one-dimensional strings of nucleotides.

Same thing, we're only beginning to explore cortex, cortex is amazingly complex and we really have very little understanding yet how complex it is. And so, I think, one should be very careful. Many people think, you know, the study of consciousness clearly can't be addressed by scientific methods and, you know, it requires extra scientific things, and I think we should just be very careful of asserting that, given that we've made this mistake several times before in our own intellectual history.

Thank you very much.

David Cohen: I'm afraid we don't have time for questions. Perhaps you can talk to Dr. Koch after the session is over.