Consciousness, gravity and the quantum: An interview with Sir Roger Penrose

Interviewer: Chris Clarke

CC. Your books have presented perhaps the most detailed proposals yet for a quantum mechanical account of consciousness, which is something that interests a lot of our readers. I'm hoping that we can explore here some of the key ideas in this work. Can we start with Quantum Mechanics? The idea that this might have something to do with consciousness isn't new, is it?

RP. No, it isn't. The trouble had been that classical physics seems rather unhelpful with regard to understanding mental qualities: there appears to be a clear dichotomy between classical science and things to do with mentality or free will. Quantum theory introduced two new features really. One was indeterminacy - so one maybe had room for free will in the indeterminacy of quantum mechanics - and the other aspect was the apparent need for an observer of some sort which sometimes people would argue had to be a conscious observer. This wasn't everybody's view, but many thought that a measurement hadn't really taken place until some conscious being had perceived it. It's one of these very unclear parts of quantum mechanics that somehow you have to have the quantum system on the one hand and the measuring apparatus, or the observer, on the other.

CC. Could you explain why we need to bring in a separate observing instrument or observer in addition to the quantum mechanical system; why couldn't one just have a purely quantum description of everything?

RP. Well, the trouble with quantum mechanics is it's a strange hybrid. Part of it is a deterministic evolution, just as happens in classical theory. This is governed by an equation first introduced by Erwin Schrödinger, so we'll call this Schrödinger evolution. Then as well as this there is the other part of quantum mechanics which is the measurement process. That takes the quantum state and does something to it which is completely different mathematically from the Schrödinger evolution, and only at that stage do you get uncertainty. There are various different "alternative realities" which are all potentially there in a quantum state. Upon measurement, only one of these alternatives becomes realised, and that realisation of a particular alternative constitutes the second quantum procedure: the measurement process.

CC. This second process is the "state reduction" that plays such a role in your work, isn't it? And I suppose the reason why you can't just have a purely Schrödinger evolution with no measurement coming in is the notorious "Schrödinger's cat" example. Can you explain that?

RP. Well, you could imagine setting up a system in which a single photon encounters what is called a beam splitter, which could simply be a half-silvered mirror, and the photon could go through it, or it could be reflected, and the two alternative things that the photon might do coexist in the quantum description. Then you imagine that the transmitted photon would trigger a device that kills a cat imprisoned in a box, while the reflected photon was harmlessly absorbed in the wall. The Schrödinger evolution part of quantum theory is linear, which means that each of the two things which might happen goes on its merry way as though the other one weren't there. They are still supposed to coexist in the quantum state, which consists of both a live cat and a dead cat. But only one of them seems to happen in actuality. You only see a cat either dead or alive. That means that something else is needed in addition to the Schrödinger evolution, and that is this strange "state reduction" that is connected with observation.

CC. So in a situation where pure Schrödinger evolution is giving you something which is an incomplete picture, something different is needed, which is achieved by adding on state reduction. But I think that as long as I've known you, you have been saying that there must be something more to quantum theory, that we actually need to transcend it, not just add things onto it.

RP. I think I've been saying that for a long time. I'm surprised how few people actually take that position. I find it quite extraordinary because it seems to be pretty obvious.

CC. So that's the quantum side of your theory. Then there is the mind side. One thing which sometimes bothers*Network* readers is, what is one meaning by mind here? Are you talking about consciousness in the sense of enjoying the sunrise and falling in love, or are you talking about problem solving, or is there something in between the two?

RP. Well, people sometimes say there are all these different forms of consciousness, and why should one lump them all together; but I would say there is one basic theme which is common to all these. Wondering at the sunrise or falling in love are certainly things in which consciousness has to feature; being in pain or perceiving a colour or hearing a melody, all these things are certainly aspects of consciousness. The only reason I haven't really addressed them in my own discussions is that I don't have anything useful to say about them. With mathematical understanding, on the other hand, which I also regard as a conscious activity, there can one begin to see that it *has to lie outside* the kinds of descriptions that people put forward, particularly the computational models. You might say that it's even less easy to see how perception of a sunrise or the feeling of pain could be the result of a computation. So if you like I'm making life hard for myself by looking at mathematics, which many people might consider to be a computational thing. It seems to me that showing that, even there, there is something which seems to transcend purely computational activity, that makes the case that much stronger.

CC. That's rather good, you've got an *a fortiori* argument. If you convince people in the case of mathematics then *a fortiori* it will bring in these other things as well. But it would probably be useful to say something about "non-computational." It's a crucial technical term that comes into the argument, isn't it?

RP. I can give examples of non-computational things, but it's not obvious why they are non-computational. The one I like the best is the tiling problem. You have these shapes made out of squares glued together, things called polyominoes. You're given a collection of these, a finite number of different sorts but an unlimited number of each sort, and you are asked: can you use these shapes to cover the entire plane without gaps or overlaps?
That is an example of a non-computational problem. That is to say, there is no computer programme which will answer yes or no for any given set of tiles: no programme where you can feed in the information of the tiles into the computer and ask it, will they tile the plane or not. Although the answer "no" can be computational, the answer "yes" is not computational. That is to say there is no way of being sure, for an arbitrary set of polyomino shapes, that they will tile the plane. It's quite a subtle piece of mathematics to show that this is a non-computational problem. There is no computer programme whatsoever that can make this decision for any possible given set of tiles.

CC. Okay, so there are some non-computational problems. Now your argument is that there are some of these problems which cannot be solved computationally, but which can be solved by bringing in human consciousness. How does this argument go?

RP. You have to phrase the problem in the right way. In the case of the tiling problem, the way it works is like this. Suppose you were given a computer programme which would answer correctly "yes" or "no" to a set of tiles, but sometimes it will say not come to any conclusion at all (you could certainly have computers like that). Then by knowing that the computer doesn't give you the wrong answer, from the computer's construction you could build a set of tiles which the computer will get stuck on: a set of tiles which you know will tile the plane, and which you also know will not be able to be answered correctly by this computer.

CC. But when you say "you know" what is it about me as a conscious human being that enables me to know, what particular faculty is it that I'm applying?

RP. It's the understanding really. It's basically Gödel's theorem, but you have to know that the computer doesn't give the wrong answers. Gödel's theorem is telling you, if you like, that the procedures we're prepared to accept as proof cannot ever be limited to specific computational procedures: they're never computationally limited because once you can phrase the rules of the computational system then you can see how to transcend them. So provided that you trust those rules, so that you're prepared to count following the rules as constituting a proof, then you can see how to get methods of proof which are outside it. Our mathematical understanding, or mathematical intuition as Gödel would put it, is something outside computation.

CC. Is the open-endedness of this procedure crucial here? I take Gödel's theorem as producing some proposition that you can't prove, but always with a bigger system in which you *can* prove it. But you don't just stay with the bigger system - isn't the crucial point here that you can always keep going out to yet another proposition that you can't prove?

RP. It is like that, that's right. And in fact in our understanding we're using this kind of procedure all the time; it's not limited to sophisticated mathematical logic. Imagine you have some procedure at which you work away, and you think you've got the rules right; and then you get worried and think maybe it's not doing everything . so you *step back* and you look at what it is that you've put into that system, and you think about what are the implications of the kinds of rules you've been using. This sort of reasoning ¾ stepping back from the system ¾ is doing the same thing as Gödel's theorem, and you always have to bring your awareness in to do that.

CC. So the insight you're talking about is self-reflexive: looking at your own thinking in order to enlarge it ?

RP. I think that's essentially right; that's basically what you're doing.

CC. So we've had quantum mechanics, and we've had consciousness As I understand it, you're saying that human beings exhibit non-computational behaviour, and then you're arguing that the only place in current physics where non-computational behaviour could come is in the reduction of the state in quantum theory. Is that how it goes?

RP. Yes: it's a Sherlock Holmes argument, "once you've eliminated the impossible then whatever remains, however unlikely, must be the truth." You wouldn't want to rely on that kind of argument if you didn't have to, but I'm afraid it's what one does have to rely on . apart from a general feeling, I suppose, based on the things we were saying about quantum mechanics when we started. So first of all you need a plausible theory, a theory which will tell you, on the physics side, at what level you expect to see state reduction objectively coming in. It would be much better if you actually had a real theory of state reduction, but I don't see that, that's for the next century maybe. At this stage the problem is just to find the kind of level at which state reduction ought to be playing a role, and trying to see anything in the brain that could be making use of this, and which could be influencing neurone activity.

CC. So whatever the theory is that you are looking for, it's got to kill two birds with one stone. It's got to carry out the state reduction and it's got to do it in a non-computational manner, so you get these human consciousness phenomena coming out.

RP. I certainly don't claim that I can see the answers to all these questions. The arguments are pretty negative in a sense, by just saying that it's hard to see how it can be anywhere else.

CC. But you are claiming the place to look for it is by bringing gravity into the picture, and this is the third impossible thing before breakfast that you are requiring us to believe!

RP. Ah, yes! And the apparent implausibility of this is that people think, quite reasonably, of gravity as something which applies to large bodies, planets going round the sun, galaxies, or maybe curved space of the universe as a whole. How is that going to be of relevance to little things going on in brains? Well, you have to go back to the quantum physics and say where would you expect to see something other than Schrödinger evolution. Let's not think about cats because that's only done for dramatic purposes and is really just complicating the issue! I'm going to think instead of moving a lump of material from one place to one slightly different place. Let's go back to our photon which encountered the beam splitter - our "half-silvered mirror" - and was put into a superposition of going one way and going another way; let's suppose that if it goes through the half-silvered mirror it encounters a device which moves this lump of material from place *A* to place *B* , whereas, if the photon were reflected, the lump is left in place *A.* Then if you follow the Schrödinger evolution of this system you come to the conclusion that this lump of material is in the superposition of being in place *A* and place *B* at the same time. That is what the quantum state is telling you.
Now the question is, accepting that it is in that superposition, at least momentarily, is it going stay that way? My position is that you have to think of that superposed state as being something like an unstable nucleus that could decay into one thing or another thing after a certain time period. The superposition has two decay modes: it could decay to *A* or it could decay to *B.* It does that in a time, according to me, which you can actually calculate, and that time is measured as the inverse of a certain energy - an energy uncertainty in the system. This way of calculating the decay time is exactly what you do with an unstable nucleus. In a uranium nucleus there is an uncertainty in its energy which is inversely related to the decay timescale, so the longer it takes to decay, the more you can reduce that energy uncertainty: it's just the Heisenberg uncertainty principle applied to energy and time.
Now in the case of the superposition of the two lumps, that energy uncertainty is calculated by looking at a conflict between the principles of quantum mechanics and the principles of general relativity. General relativity would say that each of the locations of that lump has a different space-time curved in a slightly different way, and each one of those is to be regarded as a stationary space-time in its own right, and so it has its own notion of time. The Schrödinger equation tells you the rate of change with time of the state, so you have to

know what "time" means: what does it mean to say "the rate of change with time"? Each lump location has a different idea of what "the rate of change with time" means, and to identify those two notions of time means violating the basic principle of general relativity that physics cannot depend on a particular choice of the meaning of "time". So there is a fundamental error in making that identification, an error which is basically an energy. In crude terms you can think of it as if I had my two lumps which are initially in superposition and as I pull one apart from the other one, I work out how much energy it would cost me to move away from the other considering only gravitational force. Now that energy is a very, very small energy, but if you use it in the Heisenberg uncertainty relation you find a certain timescale. And that timescale, I claim, is the decay time for that superposition: it either goes to one or the other.

CC. Isn't this "decay" actually not precisely like that of a nucleus, but something which is coming in on the outside of the conventional quantum mechanical picture?

RP. I'm using atomic decay, if you like, as an analogy and using the uncertainty principle to cover the problems one might have with energy non-conservation. It's using the reasoning, that you need that kind of flexibility if you're going to do something which looks like breaking the rules, and that flexibility is given within the Heisenberg uncertainty. You wouldn't want to do anything bigger than that, but it certainly is doing something outside standard quantum mechanics

CC. Which is precisely what we need to do, by the argument so far. Are you saying then that this is where the non-computability has to come, but you don't really know how it's going to come? Is that the situation?

RP. Yes. All that is very speculative. You can suggest models that are non-computable. I think I would look at it as something for the future, though. Maybe it can provide us with some guidance as to what kind of a theory we look for (to have it non-computable, I would argue, is something you want) but, since we don't have a theory yet, it's speculative.

CC. I find the idea of a non-computable scientific theory somehow a contradiction in terms, because the only scientific theories that I can imagine are computable ones. Can you help my failure of imagination?

RP. Well, I think you're right, the only ones we've seen have been computable, but you can certainly produce models which are completely deterministic, precisely defined, but non-computable. I've given examples in my books. Some of these are models using the polyomino tiling problem. You just take some well-defined mathematical problem such as the polyomino tile problem, and make the evolution of the universe depend on whether something will tile or not. If it will, the universe will do one thing, and if it won't it will do the other, and so it's a perfectly well defined evolution which is even deterministic, and yet there is no computer model; there's no computer simulation. So you can certainly produce model universes which are non-computable. I agree they are different from the physics models we're used to, but that doesn't make them non-scientific.

CC. I know that you've described how the quantum effects of a theory like this could be linked into the brain through these subcellular structures called microtubules, all acting in concert. But I wonder if we could look ahead, and ask where this might be heading in the future. Obviously in the long-term future we get the complete correct theory of quantum gravity with all its non-computational details and we're there! But in the short term, what are the next steps?

RP. I think there are two major strands to how one might go forward here. One is on the physics side, and there I don't see anything in what people have proposed so far on quantum-gravity schemes getting at all close to an answer to this sort of thing.

CC. They don't have this flavour of the sort of thing you're talking about.

RP. Well, hardly anyone queries the rules of quantum mechanics, they just take it all on board and don't look at trying to modify it. Some do, but they don't usually get very far. I'm not sure the other ones have got very far either, but they've got big theories! So finding the right theory is something I don't see as very immediate.
Progress is less likely to be made on the theoretical side than on the experimental, I would guess. In fact I have a proposal for an actual experiment which I'm trying to get people interested in. You do it out in space: you have to make a Schrödinger's cat out of a little crystal, a little bigger than a speck of dust, which you put into two slightly different locations about a nuclear in diameter away from each other, and you do this by sending an x-ray photon through a beam splitter. You have to keep the x-ray photon for about a tenth of a second which is the decay time for this superposed crystal and you do that by sending the photon from one satellite to another which is about an earth diameter away, and it comes back again. That's because the best way I can think of to keep an x-ray photon coherent for a tenth of a second is to reflect it between two satellites which are an earth diameter away from each other. The whole thing has to be done in space because you can't get x-rays through the atmosphere, so it's a big challenge.

CC. Could one get anywhere on the psychological side, since all this is supposed to have something to do with consciousness?

RP. Yes, well, that's the other side of the discussion. The most promising thing that I know of, again this is only an idea, was put to me by a neurophysiologist with interests in quantum mechanics called Andrew Duggins. He had a number of ideas to do with the fact that different qualities are registered in different parts of the brain, for example movement and colour. There are experiments that people have done where you investigate the way people integrate the two qualities when they look at a square that's moving and changing colour at the same time. When you have various different competing factors, how they interfere with each other might show that there's some quantum non-locality involved in that. It seems to me a really interesting set of ideas and if any experiment has an indication of a quantum role it seems this type of experiment is the most likely place to see it.

CC. So let's hope that before long more experiments will be starting to put your theory to the test. Thank you very much for sharing your ideas with our readers.

Professor Sir Roger Penrose FRS is Rouse Ball Professor of Mathematics in the University of Oxford and author of, among other books, The Emperor's New Mind, and Shadows of the Mind.

Professor Chris Clarke is Visiting Professor of Mathematics in the University of Southampton.