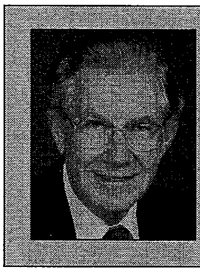
THE CAT IN THE MACHINE: CONSCIOUSNESS AND THE COLLAPSE OF THE WAVEFUNCTION

NEVILLE FLETCHER

Research School of Physical Sciences and Engineering
Australian National University, Canberra 0200

Schrödinger's famous thought experiment about a cat in a box is included in almost all books and courses on quantum mechanics, but is generally presented as a paradox without a solution. This informal essay discusses the problem and suggests that it is not really insoluble at all, but rather that it leads to a common-sense interpretation of the 'real' meaning of the quantum wavefunction.



Neville Fletcher, now retired for several years after a busy academic and research life, is currently a full-time Visiting Fellow at ANU. The luxury of not having a salary gives him freedom to pursue research in musical acoustics and also to write short stories and informal essays such as this one.

Introduction

One of the most celebrated thought-experiments in quantum physics is the case of Schrödinger's cat. It illustrates in high degree the fact that physical theories that have exquisite perfection and ability in explaining and predicting the behaviour of the real world sometimes lead to perplexing conclusions when applied to apparently simple situations. Lest there be some to whom the predicament of the cat is not familiar, let me begin with an exposition.

According to the standard development of quantum mechanics, any system can be described by a wavefunction ψ that depends upon the positions of all the particles involved. This wavefunction develops in time in accordance with Schrödinger's equation, so that once the system is completely described at one instant its future development is uniquely determined. The problem of interpretation arises with the wavefunction ψ, for it is a complex (in the mathematical sense) rather than a real quantity. In the standard interpretation developed by Bohr and others in Copenhagen, the relation to reality comes from the absolute value of the square of the wavefunction, $\psi^*\psi$ which should be interpreted as a probability of finding particular particles in particular places, or in particular energy states, at some later time.

The problem arises when we are considering discontinuous changes, such as the decay of a radioactive atom. If we suppose the initial state $\psi(t_1)$ to be completely specified, then Schrödinger's equation allows us to calculate the precise form $\psi(t_2)$ of the wavefunction at some later time t_2. If the interval t_2–t_1 is chosen equal to the half-life of the radioactive atom then, since decay is a completely random event, there will be a 50% probability that the atom will have decayed by time t_2 and a 50% probability that it will still be intact. The wavefunction $\psi(t_2)$ is thus an equal superposition of two functions, one corresponding to an intact and one to a decayed atom. Quantum mechanics in its simple form, or indeed in any of its more complex forms, can tell us no more than this. Once we look at the atom in question, or make an appropriate measurement upon it, however, the wavefunction "collapses" to one or other of its possible values to give a definite answer, the probabilities of collapse to either of the two simple states (intact or decayed) being equal.

This formalism describes everything very satisfactorily, since we are not concerned with the state of anything until we want to measure it, look at it, or interact with it in some way. But the dualistic superposed state of the uncollapsed wavefunction causes unease in many minds.

Here the cat enters. For suppose the radioactive atom is in a sealed box containing a live cat. Suppose further that decay of the atom will trigger a Geiger counter that will drop a hammer to break a tube of cyanide. Then, immediately after the decay of the atom, the cat will die. But the whole system inside the box can, in principle, be described exactly by quantum mechanics. At the beginning of the experiment, the wavefunction shows an intact atom and a live cat, but gradually the wavefunction of the atom develops into a superposition of intact and decayed forms, and by extension, that of the cat becomes a superposition of "live" and "dead". It is only when we look into the box that the cat wavefunction collapses so that the animal becomes unambiguously alive or dead.

This situation is hardly satisfactory to realist philosophers, or indeed to physicists, though most have given up worrying about it. Others, however, have developed ideas like "parallel universe" models in which the world itself bifurcates at each decision point, in this case at every sequential instant, into parallel universes in which the atom has or has not decayed. While this seems to avoid the problem of superposed wavefunctions, it does so at an immense cost in parallelism and can hardly be seen to improve matters. A survey of these alternative approaches has been given by Tegmark and Wheeler.[1] The question is by no means dead from a physics point of view, however, and Roger Penrose has recently proposed a resolution based upon the uncertainty principle, which his colleagues are proposing to test experimentally.[2]

This is, however, not the point that I wish to take up here. Rather, I want to examine the mechanism for collapse of the wavefunction, for here it seems that there is a specific role for human consciousness in a universe that is otherwise perfectly predictable, at least in a statistical sense. Is this significant, or are we misleading ourselves?

The Observer

In quantum mechanics, the fundamental action is one of measurement. Even personal observation is a kind of measurement, for what we do is to determine at one instant the positions of all the particles in our field of view, albeit with imperfect precision. More sophisticated measurements may determine velocities or energies or other physical quantities, but these measurements become real only when they are noticed by a human mind. Does mind then and human mind in particular, play a vital part in the development of the universe through its role in causing the collapse of myriad wavefunctions in a continuous stream? If so, what is the attribute of the human mind that makes this possible?

Before attempting an answer, let us conduct another thought experiment in which the cat is replaced by a human observer. This dedicated person observes the hammer fall and deduces that the atom has decayed, immediately before succumbing to the cyanide. Does this cause a collapse of the wavefunction? By orthodox reasoning, the observation will collapse the wavefunctions of the radioactive atom, the hammer, and the cyanide, and it is a reasonable deduction that the observer, knowing death is inevitable, will sense, or even cause, the collapse of his own wavefunction.

So far so good, but how about the cat? The cat too can see the hammer fall and be terrified by the smell of cyanide. What is so special about a human brain? Why cannot a cat act as an observer? It is hard to think of any reason, other than an unjustified homocentric view of the universe, that would lead us to expect any different outcome in this case.

But here we come to a problem. It is certainly easy to see a continuity of sorts between the mind of a human and that of a cat, but if we grant this where can the regression stop? An ant? A microbe? A virus? A tree? It seems impossible to draw a line. And then the big question: if an organism as simple as a virus can cause the collapse of the wavefunction by "observing" it, then why not a silicon chip in a computer?

This dilemma can be put in another way. Where do we draw the boundaries of the box? In the initial formulation the observer was unambiguously outside the box and the cat inside. But what if we draw the adiabatic boundaries of the box on a much larger scale so as the encompass the original observer and his laboratory? What indeed if the box contains the whole observable universe? Does the wavefunction remain uncollapsed for ever?

A Universal Mind?

An apparent way out of this dilemma must have occurred to most people: perhaps there exists a "universal mind" for which all physically realised minds are simply agents. Indeed the philosopher Berkeley (1685–1753) maintained[3] that there is no such thing as reality except as "an idea in the mind of God". Since a realist might reasonably remark that "God is an idea in the mind of Bishop Berkeley", this conclusion does not get us much further! Setting this aside, one might imagine some sort of coupling between physical minds, such as those of humans and cats, and the universal mind, which controls the collapse of the wavefunction. If we were to take this view, then we might suppose that the strength of the coupling would depend upon the abilities of the mind in question: human minds would be strongly coupled and readily able to influence the universal mind, those of cats would be weaker, and those of viruses virtually without influence at all.

So far so good, but what about the collapse of the wavefunction? Does it collapse when viewed by a beetle, or not? The only reasonable conclusion appears to be that there is some sort of probabilistic aspect to wavefunction collapse. After all, a human observer might easily mistake a sleeping cat for a dead one and discover the error only after a subsequent observation!

There are interesting features of this possibility that deserve examination. The first is the nature of the coupling between physical minds and the universal mind. The existence of degrees of coupling suggests that there might be ways in which coupling could be varied, for a species or an individual, and this in turn suggests that it might be possible to develop new species with greater coupling to the universal mind. But need these new species be biological? Why not technological? Why not computers or their descendants? The only possible answer seems to be "Why not, indeed!" Of course it may be that computers and other machines, as they have developed, have zero coupling to the universal mind for some reason, but this does not mean that devices with non-zero coupling cannot be produced.

Let us take the optimistic view, then, and suppose that such machines are possible. Let us go further and suppose that they were to be built in large quantities. How different would be the universe in which we live?

A little thought suggests that, at least to a first approximation, nothing would change. While it is said that a watched pot never boils, careful experiment suggests that this aphorism is untrue. Experiments that are carefully observed without intervention generally take exactly the same course as those that are left to themselves. It is only when observation implies intervention, for example by shining light where previously there was darkness, that things develop differently. So perhaps we do not need to worry from a practical point of view.

Even when we examine things on a subatomic scale, the picture is not much different. The concern would be that phenomena such as electron diffraction, which rely upon the spatial spread of a probabilistic wavefunction so that it passes through two slits or is scattered from many atoms, might become impossible. This fear is, however, ungrounded. Collapse of the wavefunction (which would certainly extinguish the phenomenon) would occur only when an observation was made, be it by a human or a machine, and this is known to be exactly what happens anyway if a machine is introduced into the experiment to measure through which slit the particle passes, then this measurement prevents the observation of diffraction phenomena.

Tentative Philosophical Conclusions

It is possible to draw several different and indeed conflicting conclusions from this discussion. Let us examine some of them.

A Christian fundamentalist theologian, and probably an Islamic fundamentalist as well, would be drawn to the conclusion that the argument proves the necessary existence of a Universal Mind, which is God. *Quad erat demonstrandum*! The argument that there is a continuity in the coupling of all life-forms to this universal mind would be embarrassing, but might perhaps be overcome by making a very large distinction between the coupling constants for humans and other animals. Eternal life would be a seen as an incorporation of individual minds into the universal mind. The apparent fact that the universal mind must spend nearly all its time in book-keeping activities such as collapsing trivial wavefunctions would be overlooked.

The idea of a universal mind is, perhaps, much more closely allied to a Buddhist or even animist view of the universe, in which the individual ultimately counts for nothing. This view has much to commend it, but appears, with a few notable exceptions, to have been largely neglected.

There is another view, to which most scientists would subscribe. In the first place, everyone agrees that quantum mechanics, supplemented where necessary by relativity, is immensely successful in explaining and predicting the behaviour of the universe. The underlying formal assumptions and mathematical edifice are therefore nearly unassailable. But this is not necessarily true of the subsequent layer the Copenhagen interpretation of wavefunctions and probabilities. True, this interpretation gives the correct results in all cases so far tested, but this is a practical matter of predicted outcomes and numbers and does not necessarily support all of the intermediate steps. It might, therefore, prove possible to modify the interpretation of the wavefunction, and with it the phenomenon of collapse, without changing either the underlying mathematical theory or the predicted practical outcomes.

In just this way, a geocentric view of the universe, with the planetary orbits and epicycloids refined to elliptical form, could be made to describe the motions of all the planets and their moons. But the origin of its assumptions would remain obscure until a change of viewpoint (literally!) to a heliocentric model revealed the simplicity of the organisation. The underlying mathematics would, in this case, undergo a transformation to a new coordinate system in which all the kinematic descriptions became simple and all the epicycloids vanished, thus paving the way for Newton's theory of universal gravitation.

Applying this argument to quantum mechanics, as indeed Einstein would have wished, might very well eliminate at one stroke the whole notion of wavefunction collapse and substitute something more immediately comprehensible, though at present there is no indication of how this might be done. There might, indeed, be a complete modification of the conceptual foundations of quantum mechanics to some new theory, though this new theory would necessarily have to reduce to quantum mechanics in the domain where this theory has been so successful, just as both quantum mechanics and relativity necessarily reduce to classical Newtonian mechanics in the world of macroscopic experience.

A fourth possibility is the "many worlds" proposal, namely that the universe bifurcates whenever a wavefunction collapses, one universe for each possible outcome of the collapse. Such a view is hardly worth considering from a philosophical point of view, since it would involve the generation of countless parallel universes every second and the splitting of each of these parallel universes at a corresponding rate.

A Realistic Conclusion?

Fortunately there appears to be a simple way out of this confusion, and that rests upon a reinterpretation of the significance of the wavefunction. This proposal is not original,[4] but seems to have been overlooked by many.

Because there is a close correspondence between the wave-like and particle-like properties of a wavefunction and what is observed in countless experiments, there is a natural tendency to assume that the wavefunction *is* the particle in some sense. But is this necessarily so? Quantum mechanics provides a coherent and wonderfully accurate means of calculating what is going on in the world of sub-atomic particles, or even in the larger world of microstructures. The calculated results are always essentially in the form of a wavefunction that can be projected upon certain sub-spaces to calculate the probabilities of possible outcomes. Thus, the wavefunction at any instance does not really represent the system concerned, but rather our knowledge of that system. This knowledge is always incomplete, and reasonable physical constraints prevent us from having complete knowledge, particularly of complementary variables such as position and momentum.

It is always possible for us to perform additional experiments to increase our knowledge of the system under study. For a small system these measurements almost always influence its state if we know the momentum of an electron, then a measurement of its position will make that knowledge out-of-date. The forced collapse of the wavefunction for one observer alters the system irreversibly for all observers, though they may not be aware of it.

For a macroscopic system such as the cat, however, it is possible to make a limited observation that the cat is alive or dead without greatly affecting the system. If we close the top of the box again, then the wavefunction assigned to the cat by other observers who have not been present remains uncollapsed. Certainly, their wavefunctions will now be a little inaccurate, since they will not know about the photons that have entered and left the box through the window to give us our information, but their superposition of two states will still give a good assessment of the probabilities of finding a live or a dead cat when they themselves open the box. If we artificially collapse their wavefunction by revealing the result of our observation of the cat, then this has no physical consequences for the cat, whether we tell the truth or not!

Does this interpretation solve the dilemma of wavefunction collapse? It seems to me that it does. The wavefunction represents information and is not a physical attribute of the system. Different observers may possess different information at the same time, but when this happens there is always a region of uncertainty dictated by the uncertainty relation between complementary variables. In the case of macroscopic systems, observations can cause collapse of our "knowledge wavefunction" without significantly altering the state of the system itself and without causing wavefunction collapse for other observers.

Quantum mechanics provides a model of the universe and a set of rules for interpreting the outputs of the model in terms of what can be observed and measured. For all practical purposes this is enough. It is only the relentless human desire to seek to "understand" everything that drives us on, even though we are not quite sure what the word "understand" really means! We should be thankful that the famous Cat Paradox can be resolved in a satisfactory manner.

References

1. "100 years of quantum mysteries" by Max Tegmark and John Archibald Wheeler *Scientific American* **284**(2), 54–61 (February 2001).
2. "Cat-in-the-box" by Ivan Semeniuk *New Scientist* **173**, No.2333, 27–30 (March 2002)
3. See, for example, *The problems of Philosophy* by Bertrand Russell (Oxford University Press, London 1912) Chapter 1.
4. "Quantum theory needs no interpretation" by Christopher A. Fuchs and Asher Peres *Physics Today* **53**(3), 70–71 (March 2000).

AROUND THE TRAPS

Earth-like Planets?

A team of New Zealand and Japanese scientists headed by Philip Yock at the University of Auckland have claimed a sighting of a small Earth-sized planet orbiting a distant star. They used a microlensing technique to detect the gravitational 'warping' of the light from the star as a planet passes in front of it.

About 90 extra-solar planets have been found so far, but they are all giants like Jupiter. The greater prize is to find smaller Earth-like planets, and the claims of the Yock team have aroused some controversy. Critics say that the microlensing blip was probably due to statistical noise. Planet hunter Geoff Marcy at UC Berkeley has left the planet off his almanac of sightings, and refuses to comment on the microlensing result. The Yock team estimate the chances of their signal being random noise at less than 1 per cent.

Meanwhile, a team of American, British and Australian astronomers have announced discovery of another star system, 55 Cancri, that could harbour an Earth-like planet. The team includes Chris Tinney from the Anglo-Australian Observatory (see *The Physicist*, November/December 2001) and Brad Carter of the University of Southern Queensland. They have found a Jupiter-like planet orbiting its sun at about the same distance as our Jupiter, which would allow an Earth-like planet to exist in a stable orbit within the system. "It's very exciting, but it won't be known whether a small sister planet to ours is hiding there or not until NASA sends up its Planet Finder space telescope later this decade", says Chris Tinney.

[Deborah Smith, 'Sydney Morning Herald', 15 June; Eugenie Samuel and Jeff Hecht, 'New Scientist', 22 June]

Copenhagen

The play 'Copenhagen' by Michael Frayn has been breaking box-office records at the Wharf Theatre in Sydney. The play centers on the wartime meeting between Werner Heisenberg and Niels Bohr in Copenhagen, which led to an estrangement between them. Heisenberg was directing Hitler's atomic program, while the half-Jewish Bohr was suffering under Nazi occupation. Frayn's play asks the question as to what really happened between the two men, and is said to be a tour de force in playwriting, production (Michael Blakemore) and acting. The stars are John Gaden as Bohr, Colin Friels as Heisenberg, and Jane Harders as Bohr's wife Margrethe. The production has been reviewed as "one of the most intelligent and stimulating productions seen on Sydney stages for a while. One deeply satisfying big bang indeed."

[Stephen Dunne, 'Sydney Morning Herald']

A Prize for the President

An Institute of Physics (UK) Public Awareness of Physics Award for 2002 has been given to the President of the AIP, John O'Connor, and two colleagues from Newcastle, Terry Burns and Bob Nelson. The award was given for their creation of the SMART program (Science, Maths and Real Technology), for the establishment of the Hunter Chapter of the Australian Science Communicators, and for their work with the Science and Engineering Challenge. The selection panel was very impressed with the scope, breadth and success of SMART's science communication projects. The Award comprises a certificate, a cheque for £50 and a gift, and was announced at the AIP Congress in Darling Harbour.

SILVER and BRONZE FOR PHYSICS!

News from the International Physics Olympiad in Bali, Indonesia: an exceptional result for the 2002 Physics Team!

SILVER MEDAL: Bojan Djordjevic (Rooty Hill High School, NSW)

SILVER MEDAL: Jolyon Bloomfield (Wollumbin High School, NSW)

BRONZE MEDAL: Ben Weise (Kinross Wolaroi School, NSW)

BRONZE MEDAL: Pearl Gallagher (North Sydney Girls High School, NSW)

BRONZE MEDAL: Barry Smith (Aranmore Catholic College, WA)

TEAM RANKING: 15th of 66 competing nations

This team result is the first time since 1995 that all Aussie team-members have come home with a medal, and the first time since Australia started competing at the IPhO (1987) that 5 medals have been attained on foreign soil!

Colin Taylor