Quantum Mechanics and Experience

• • • • • •

David Z Albert

Harvard University Press
• • •
Cambridge, Massachusetts
London, England

experimental evidence that the quantum state of any isolated physical system in the world ever fails to evolve in perfect accordance with the linear dynamical equations of motion.[29]

And so there seem to be a number of good reasons for looking for a different angle on this whole business.

[29] That is: there hasn't been a shred of evidence that such failures ever take place, aside from the outcomes of certain *introspective* experiments that we carry out on *ourselves.*

••• 6 •••

The Dynamics by Itself

What It Feels Like to Be in a Superposition

The trouble with the quantum-mechanical equations of motion, according to Chapter 4 (and according to the conventional wisdom), runs as follows: The equations of motion (if they apply to everything) entail that in the event that somebody measures (say) the color of a hard electron, then the state of the measured electron (e) and the measuring device (m) and the human experimenter (h), when the experiment is over, will be:

$$1)\qquad 1/\sqrt{2}(|\text{believes } e \text{ black}\rangle_h|\text{“black”}\rangle_m|\text{black}\rangle_e + |\text{believes } e \text{ white}\rangle_h|\text{“white”}\rangle_m|\text{white}\rangle_e)$$

and of course, the state in (6.1) is (on the standard way of thinking) a state in which there is no matter of fact about what the color of the electron is, or about what the measuring device indicates its color to be, or even about what the experimenter takes its color to be; and the trouble with *that* (according to Chapter 4) is that we know, with certainty, by means of direct introspection, that there *is* a matter of fact about what we take the color of an electron like that to be, once we're all done measuring it; and so the state in (6.1) can't possibly be the way that experiments like that end up; and so the linear equations of motion must (in some instances, at least) be false.

But there's a small tradition of *resistance* to that conventional wisdom, which goes back to the late Hugh Everett III, who announced, in 1957, in a paper which is both extraordinarily suggestive and extraordinarily hard to understand, that he had discovered a way of coherently entertaining the possibility that the linear quantum-mechanical equations of motion are indeed (notwithstanding the argument rehearsed above) the true and complete equations of motion of the whole world. And that tradition merits some of our attention here.

What Everett announced (to put it a little more concretely) was that he had discovered some means of coherently entertaining the possibility that the states of things at the conclusions of color measurements of initially hard electrons really *are* (precisely as the linear equations of motion demand) superpositions like the one in (6.1); and the idea of what has become the canonical *interpretation* of Everett's paper (see, for example, DeWitt, 1970) is that the means of coherently entertaining that possibility that Everett must have had in mind (or perhaps the one which he *ought* to have had in mind) is to take the two components of a state like the one in (6.1) to represent (literally!) two physical *worlds.* The idea is that in the course of a measurement of (say) the color of a hard electron (the sort of measurement, that is, that leads to the state in (6.1)) the number of physical worlds there are literally increases from one to two, and that in each one of those worlds that color measurement actually has an outcome and the observer actually has a determinate belief about that outcome, and that those worlds are subsequently absolutely unaware of one another.

That's interesting. But there's a sense in which it can't be the whole story; there's a sense (that is) in which the above sort of talk, as it stands, isn't *well defined.* The trouble is that what worlds there are, at any particular instant, on this way of talking, will depend on *what separate terms there are in the universal state vector* at that instant; and what separate terms there are in that state vector at that particular instant will depend on what *basis* we choose to *write that vector down in;*[1] and of course there isn't anything in the

[1] Here's what I mean: Suppose that a state like the one in (6.1) obtains; and consider (on this way of talking) what *worlds* that means that there presently are. What we're *tempted* to say (under these circumstances, on this way of talking) is that when (6.1) obtains there's one world in which the electron is black and there's

quantum-mechanical formalism itself which will pick any particular such basis out as the (somehow) *right* one to write things down in;[2] and so, if there's going to be any objective matter of fact about what worlds there are, at any given instant, on the many-worlds way of talking, then some new general principle is going to have to be *added* to the formalism which *does* pick out some particular basis as the right one to write things down in.

And of course the kind of principle we'll need is one that can guarantee that the worlds that come into being in the course of anything that can count as a measurement are all worlds in which there's a matter of fact about how that measurement *comes out.*

And so figuring out exactly what that principle ought to be is going to amount to figuring out exactly which *physical variables* there need to be matters of fact about, in order for there to be matters of fact about how measurements come out. And of course the business of figuring *that* out (as we discovered in the course of trying to cook up postulates of collapse) turns out to be very difficult.

And anyway, there's a more serious problem. There's a puzzle, when you talk like this, about what it could possibly mean to say (for example) that in the event that *h* carries out a color measure-

another world in which it's white. But note that the state in (6.1) could *also* have been written down like this:

$$1/\sqrt{2}(|Q+\rangle_{h+m}|\text{hard}\rangle_e + |Q-\rangle_{h+m}|\text{soft}\rangle_e)$$

where

$$|Q+\rangle_{h+m} = 1/\sqrt{2}(|\text{believes } e \text{ black}\rangle_h|\text{``black''}\rangle_m + |\text{believes } e \text{ white}\rangle_h|\text{``white''}\rangle_m)$$

and

$$|Q-\rangle_{h+m} = 1/\sqrt{2}(|\text{believes } e \text{ black}\rangle_h|\text{``black''}\rangle_m - |\text{believes } e \text{ white}\rangle_h|\text{``white''}\rangle_m)$$

and *that* makes things look (on this way of talking) as if there's one world in which the electron is *hard* and there's another in which it's *soft!*

[2] On the standard quantum-mechanical formalism, after all, the choice of a set of basis vectors in which to write states down has no *physical* significance *whatso-*

ment of a hard electron, the "probability" that that measurement will come out white is ½. The trouble is that that sort of a measurement (on *this* way of talking) will *with certainty* give rise to *two worlds,* in one of which there's an *h* who sees that the outcome of the measurement is white, and in the other of which there's an *h* who sees that the outcome of the measurement is black; and there isn't going to be any matter of fact about which one of those two worlds is the *real* one, or about which one of those two *h*'s is the *original h.*

And there are myriad other difficulties with talking the many-worlds talk too (see, for example, Barrett, 1992), but I guess we need not rehearse any more of them here.

There are ways of making many-worlds talk sound less vulgar (which is to say: there are ways of making it sound less literal). But they don't get at what the real problems are.

Sometimes it gets proposed (for example) that there is exactly one physical world but that (when states like (6.1) obtain) there are two incompatible *stories* about that world, or maybe about how *h sees* that world, which are both somehow simultaneously *true.*[3]

It seems to me that that's really hard to understand. But one of the things that's obvious about it is that it runs into exactly the same sort of puzzle about what *probabilities* mean as the many-*worlds* talk does. Suppose, for example, that an observer named *h* carries out a measurement (just like the one we talked about above) of the color of a hard electron. Try to figure out what it might mean to say of an experiment like that (if you try to talk like this) that the probability that its outcome will be black is ½. The trouble is that this sort of talk is going to entail, with certainty, that there are *two* stories about what happens in an experiment like that; and there isn't going to be any matter of fact about *which one* of those

[3] The most interesting attempt I know of at talking like that is Michael Lockwood's, in *Mind, Brain, and the Quantum* (Lockwood, 1989). Lockwood tries harder than anybody else does (with the possible exception of Lockwood's colleague David Deutsch, whose ideas show up at a number of crucial points in *Mind, Brain, and the Quantum*) to think about what it means (that is: to think about what it's *like*) for there to be more than one true story, when a state like (6.1) obtains, about what *h*'s experience is.

stories is the *true* one, and there isn't going to be any matter of fact about which one of those stories is the one that's about *the original h.*

I think it turns out to be a good deal more interesting to read Everett in a rather different way.

Suppose that there is only one world, and suppose that there is only one full story about that world that's true, and suppose that the linear quantum-mechanical equations of motion are the true and complete equations of motion of the world, and suppose that the standard way of thinking about what is means to be in a superposition is the right way of thinking about what it means to be in a superposition, and consider the question of what it would *feel* like to *be* in a state like the one in (6.1) (that is: the question of what it would feel like to be *the experimenter* in a state like the one in (6.1)).

That question wasn't confronted in Chapter 4. There didn't seem to be much of a point (back then) in confronting it. What seemed important was just that whatever it might feel like to be in a state like the one in (6.1), it certainly would *not* feel like what *we* feel like when we're all done measuring the color of a hard electron.[4]

But (since it turns out not to be easy to cook up a good-looking theory of the collapse, and since it turns out that no theory of the collapse whatsoever is going to be able to preclude the development of states like the one in (6.1) in people who undergo the kind of brain surgery described at the end of Chapter 5, and since there aren't any normal experimental reasons for believing that there *are* any such things as collapses) things are different now.

Here's a way to get started:

Suppose that the linear quantum-mechanical equations of motion were invariably true and (consequently) that observers like the one described above frequently did end up, at the conclusions of color measurements, in states like the one in (6.1).

Let's see if we can figure out what those equations would entail

[4] That is: what seemed important was just that it could be established (by means of the argument on page 112) that as a matter of fact human experimenters don't *end up* in states like that, at the conclusions of those sorts of measurements.

about how an observer like that, in a state like the one in (6.1), would *respond* to *questions* about how she feels (that is: about what her mental state is). Maybe that will tell us something.

The most obvious question to ask is: "What is your present belief about the color of the electron?" But that question turns out not to be of much use here. Here's why: Suppose that the observer in question (the one that's now in the state in (6.1)) gives honest responses to such questions; suppose, that is, that when her brain state is $|$believes *e* black$\rangle$ she invariably responds to such a question by saying the word "black," and when her brain state is $|$believes *e* white$\rangle$ she invariably responds to such a question by saying the word "white." The problem is that precisely the same linearity of the equations of motion which brought about the superposition of different brain states in the state in (6.1) in the first place will now entail that if we were to address this sort of a question to this sort of an observer, when (6.1) obtains, then the state of the world after she *responds* to the question will be a superposition of one in which she says "black" and another in which she says "white"; and of course it won't be any easier to interpret a "response" like that than it was to interpret the superposition of brain states in (6.1) that that response was intended to be a *description* of!

But there are other sorts of questions that turn out to be more informative.

Note, to begin with, that it follows from the linearity of the operators that represent observables of quantum-mechanical systems (the sort of linearity that was defined in equation (2.9)) that if any observable O of any quantum-mechanical system S has some particular determinate value in the state $|A\rangle_S$, and if O also has that same determinate value in some other state $|B\rangle_S$, then O will necessarily *also* have precisely that same determinate value in any linear *superposition* of those two states.[5]

[5] That's an entirely commonsensical way for observables to behave, if you think it through. Suppose, for example, that there's a particle which is in a superposition of being located in the right half and in the left half of a certain box. What the linearity of the observables of a particle like that is going to entail (or rather, *one* of the things that it's going to entail) is that that particle is in an eigenstate of the observable "is the particle anywhere in the box at all?" with eigenvalue "yes."

Let's apply that to the superposition of states in (6.1).

Suppose that we were to say this to *h*: "Don't tell me whether you believe the electron to be black or you believe it to be white, but tell me merely whether or not *one* of those two is the case; tell me (in other words) merely whether or not you now *have* any particular definite belief (not uncertain and not confused and not vague and not superposed) about the value of the color of this electron."

Now, if (when we ask *h* that) the state $|\text{believes } e \text{ black}\rangle_h \times |\text{"black"}\rangle_m|\text{black}\rangle_e$ obtains, and if *h* is indeed an honest and competent reporter of her mental states, then she will presumably answer, "Yes, I *have* some definite belief at present, one of those two *is* the case"; and of course she will answer in precisely the same way in the event that $|\text{believes } e \text{ white}\rangle_h|\text{"white"}\rangle_m|\text{white}\rangle_e$ obtains.

And so responding to this particular question in this particular way (by saying "yes") is an observable property of *h* in both of those states, and consequently (and this is the punch line) it will also be an observable property of her in any superposition of those two brain states, and consequently (in particular) it will be an observable property of her in (6.1).

That's odd. Look what we've found out: On the one hand, the dynamical equations of motion predict that *h* is going to end up, at the conclusion of a measurement like the one we've been talking about, in the state in (6.1), and not in either one of the brain states associated with any definite particular belief about the color of the electron; on the other hand, we have just now discovered that those same equations also predict that when a state like (6.1) obtains, *h* is necessarily going to be convinced (or at any rate she is necessarily going to *report*) that she *does* have a definite particular belief about the color of the electron. And so when a state like (6.1) obtains, *h* is apparently going to be radically deceived even about what *her own occurrent mental state* is.

And so it turns out that there was a hell of a lot too much being taken for granted when we got convinced (back in Chapter 4) that there is some particular point in the course of the sort of measurement we've been talking about by which a collapse of the wave function must necessarily already have taken place, some particular point (that is) at which the dynamical equations of motion together with the standard way of thinking about what it means to be in a superposition somehow flatly contradicts what we unmistakably know to be true of our own mental lives.

Let's go on. Suppose that *h* carries out a measurement of the color of a hard of electron with a color measuring device called *m*1, and suppose that when that's done (that is: when a state like (6.1) obtains) *h* carries out a *second* measurement of the color of that electron, with a *second* color measuring device called *m*2. When *that's* all done, the state of *h* and of the two measuring devices and the electron (if the measuring devices are good, and if *h* is competent, and if everything evolves in accordance with the linear dynamical equations of motion) is going to look like this:

2) $1/\sqrt{2}(|$believes outcome of first measurement is "black"
and believes outcome of second measurement is
"black"$\rangle_h|$"black"$\rangle_{m1}|$"black"$\rangle_{m2}|$black$\rangle_e$
$+\ |$believes outcome of first measurement is "white"
and believes outcome of second measurement is
"white"$\rangle_h|$"white"$\rangle_{m1}|$"white"$\rangle_{m2}|$white$\rangle_e)$

And suppose that at that point (when (6.2) obtains) we were to say to *h*: "Don't tell me what the outcomes of either of those two color measurements were; just tell me whether or not you now believe that those two measurements both had definite outcomes, and whether or not those two outcomes were *the same.*"

It will follow from the same sorts of arguments as we gave above that *h*'s response to a question like that (even though, as a matter of fact, on the standard way of thinking, *neither* of those experiments had any definite outcome) will necessarily be: "Yes, they both had definite outcomes, and both of those outcomes were the same."

And it will also follow from the same sorts of arguments that if two observers were both to carry out measurements of the color of some particular initially hard electron, and if they were subsequently to talk to one another about the outcomes of their respective experiments (if they were both, that is, to *check up* on one another), then both of those observers will report, falsely, that the *other* observer has reported some definite particular outcome of *her* measurement, and both of them will report that that reported outcome is completely in agreement with her own.

Let's make up a name for all that. Let's say that when a state like (6.1) obtains, then (even though there isn't any matter of fact about what the color of the electron is, and even though there isn't even any matter of fact about what *h*'s *belief* about the color of the electron is) what the dynamics entails is that *h* "effectively knows" what the color of the electron is.

Let's go on some more. Suppose that *h* is confronted with an infinite collection of electrons, all of which are initially hard, and that *h* undertakes to measure the color of *each one* of those electrons.

Before those measurements start, the state of *h* and of those electrons (whose names are 1, 2, . . .) and of *h*'s color measuring devices (whose names are, respectively, *m*1, *m*2, . . .) is:

3)

$$|\text{ready}\rangle_h|\text{ready}\rangle_{m1}|\text{hard}\rangle_1|\text{ready}\rangle_{m2}|\text{hard}\rangle_2|\text{ready}\rangle_{m3}|\text{hard}\rangle_3\ \ldots$$

Once the measurement of the color of electron 1 is done, the state is:

4)

$$\begin{aligned}&1/\sqrt{2}(|\text{believes 1 black}\rangle_h|\text{“black”}\rangle_{m1}|\text{black}\rangle_1\\ &\quad+|\text{believes 1 white}\rangle_h|\text{“white”}\rangle_{m1}|\text{white}\rangle_1)\\ &\times|\text{ready}\rangle_{m2}|\text{hard}\rangle_2|\text{ready}\rangle_{m3}|\text{hard}\rangle_3\ \ldots\end{aligned}$$

And once the measurement of the color of electron 2 is done, the state is:

5)

$$\begin{aligned}&1/\sqrt{4}\{(|\text{believes 1 black and 2 black}\rangle_h|\text{“black”}\rangle_{m1}\times\\ &\qquad|\text{“black”}\rangle_{m2}|\text{black}\rangle_1|\text{black}\rangle_2)\\ &\quad+(|\text{believes 1 black and 2 white}\rangle_h|\text{“black”}\rangle_{m1}\times\\ &\qquad|\text{“white”}\rangle_{m2}|\text{black}\rangle_1|\text{white}\rangle_2)\\ &\quad+(|\text{believes 1 white and 2 black}\rangle_h|\text{“white”}\rangle_{m1}\times\\ &\qquad|\text{“black”}\rangle_{m2}|\text{white}\rangle_1|\text{black}\rangle_2)\end{aligned}$$

$$+ \; (|\text{believes 1 white and 2 white}\rangle_h|\text{“white”}\rangle_{m1} \times$$
$$|\text{“white”}\rangle_{m2}|\text{white}\rangle_1|\text{white}\rangle_2)\}$$
$$\times \; |\text{ready}\rangle_{m3}|\text{hard}\rangle_3 \; . \; . \; .$$

And so on. The number of separate mathematical terms in the state vector of the world (if you write it out in the sort of basis that's used here) will increase geometrically (like the numbers of the branches in the diagram in figure 6.1, as you work your way up) as the number of color measurements increases.

Now, suppose that once the first N of those measurements are complete we say this to h: "Don't tell me what the color of electron 1 or electron 2 or any particular one of the first N electrons turned out to be; tell me merely whether or not you believe that each one of those electrons now has a definite color, and tell me also (if the answer to that first question is yes) what *fraction* of those first N electrons turned out to be *black*."

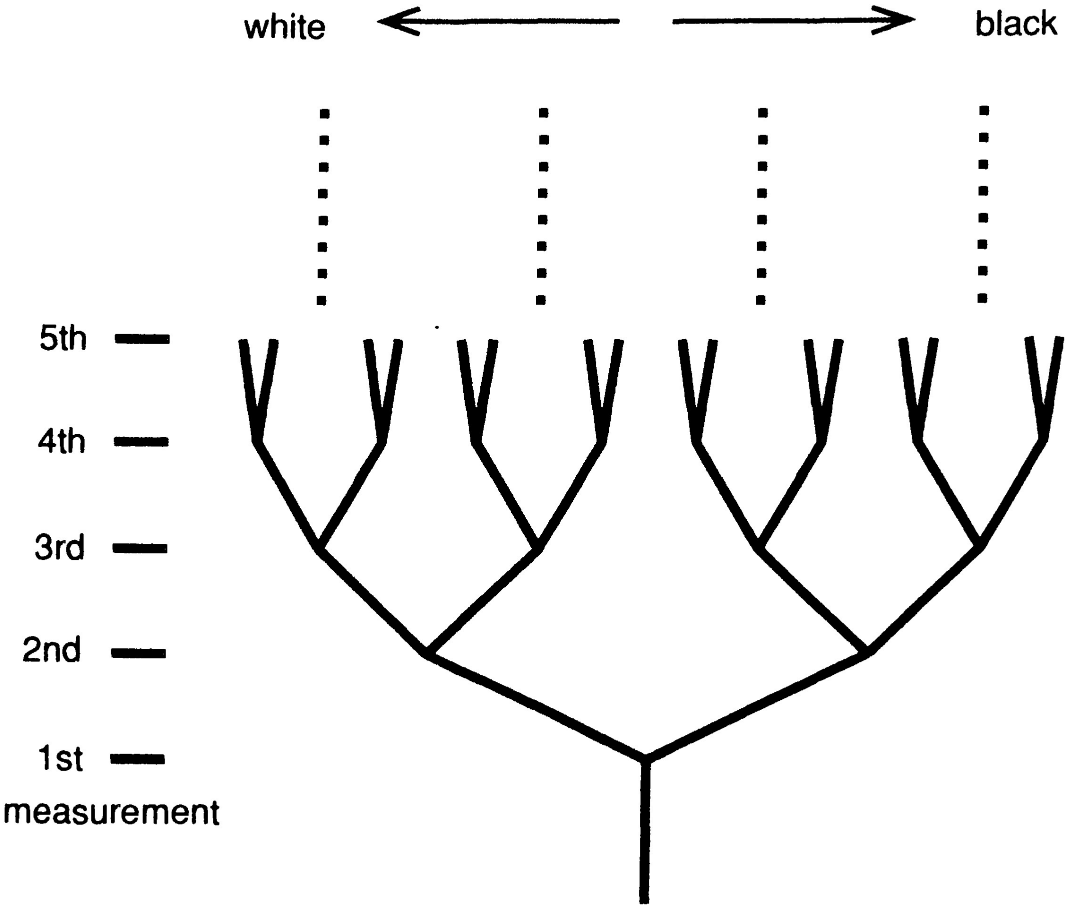

Figure 6.1

That won't tell us much, as it stands. The answer to the first question (as we've already seen) is going to be "yes" (and moreover, at that point, it's going to be a physical fact about the world that *h effectively knows* the color of each of those first *N* electrons). But of course *h isn't* going to produce any coherent answer to the *second* question; once *h* has responded to *that* question (if *h* is a competent converser on these matters), the state of the world is going to be a superposition of states (like the superposition that arises in the event that we ask *h* what the color of the electron is when (6.1) obtains) in which *h* answers that question in various different ways.

But here's something curious: It happens that in the limit as *N* goes to infinity (that is: in the limit as the number of color measurements which *h* has so far performed goes to infinity), the state of the world will, with certainty, *approach* a state in which *h will* answer that question in a perfectly *determinate* way, and in which the answer *h* gives will with certainty be "½" (which is, of course, precisely what *ordinary* quantum-mechanics will predict, with certainty, about that response, in that limit).[6]

And that turns out to be an instance of something a good deal

[6] It isn't hard to see why that sort of thing ought to be true. Here's how to start out: Consider (for example) an infinite collection of electrons (let's call them 1, 2, 3, . . .), all of which are in the |hard⟩ state; and consider the following *observable* of that collection: $O_N = (1/N) \times$ (the number of *black* electrons among the first N electrons).

Now, *ordinary* quantum mechanics (that is: quantum mechanics with a *collapse*) entails that if the color of each one of those electrons were to be measured, then (since there are infinitely many of those electrons in this collection) precisely half of those measurements would with certainty come out "black," and precisely half of them would with certainty come out "white." Moreover, O_N (for any value of N) is *compatible* with the colors of every one of those electrons. And so it follows that ordinary quantum mechanics entails that as N approaches infinity, the probability that a measurement of the value of O_N on the collection of electrons described above will find the value ½ will approach 1. And so it follows that that collection of electrons (*prior* to any measurement) must be in an *eigenstate* of whatever operator it is that O_N approaches as *N* approaches infinity, with eigenvalue ½.

And it will follow from all *that* that *whether or not* there are ever any such things as collapses, the state of a composite system consisting of that collection of initially more general, which runs as follows: Suppose that an observer *h* is confronted with an infinite ensemble of identical systems in identical states and that she carries out a certain identical measurement on each one of them. Then, even though there will actually be no matter of fact about what *h* takes the outcomes of *any* of those measurements to be, nonetheless as the number of those measurements which have already been carried out goes to infinity, the state of the world will approach (as a well-defined mathematical limit) a state in which the reports of *h* about the *statistical frequency* of any particular outcome of those measurements will be perfectly definite, and also perfectly in accord with the standard *quantum-mechanical* predictions about what that frequency ought to be.

So it turns out not to be altogether impossible (even if the standard way of thinking about what it means to be in a superposition is the right way of thinking about it) that the state we end up in at the conclusion of a measurement of the color of a hard electron is the one in (6.1). And so everything we've been thinking about the measurement problem up till now isn't right.

And what all this obviously suggests is that maybe there just isn't any such thing as a measurement problem.

That is: maybe (even if the standard way of thinking about what it means to be in a superposition is the right way of thinking about what it means to be in a superposition) the linear dynamical laws are nonetheless the complete laws of the evolution of the *entire world,* and maybe all of the appearances to the contrary (like the appearance that experiments have outcomes, and the appearance that the world doesn't evolve deterministically) turn out to be just the sorts of *delusions* which *those laws themselves* can be shown to *bring on!*

hard electrons and of a competent *observer* who has just carried out measurements of the *colors* of all those electrons will, with certainty, be an eigenstate of that observer's *reporting* (if she's asked) that the value of whatever operator it is that O_N approaches as N approaches infinity is ½.

A detailed mathematical discussion of all this stuff (with much nicer proofs than the one above) can be found in the doctoral dissertation of Jeff Barrett (1991).

This is an amazingly cool idea (let's call it "the bare theory"), and *this* is the idea that it strikes me as interesting to read into Everett's paper.[7]

Notwithstanding all the stuff we've just learned, however, it seems to me that the bare theory can't be quite right either.

Note, for example, that if the bare theory is true, then there will be matters of fact about what we think about (say) the frequencies of "black" outcomes of measurements of the color of hard electrons *only* (if at all) in the limit as the number of those measurements goes to infinity. And so, if the bare theory is true (and since only a finite number of such measurements has ever actually been carried out by any one of us, or even in the entire history of the world), then there can't now be any matter of fact (notwithstanding our delusion that there is one) about what we take those frequencies to *be.* And so, if the bare theory is true, then there can't be any matter of fact (notwithstanding our delusion that there is one) about whether or not we take those frequencies to be in accordance with the standard quantum-mechanical *predictions* about them. And so, if the bare theory is true, it isn't clear what sorts of reasons we can possibly have for *believing* in anything like quantum mechanics (which is what the bare theory is supposed to be a way of making *sense* of) in the *first* place.[8]

And as a matter of fact, if the bare theory is true, then it seems extraordinarily unlikely that the present quantum state of the world can possibly be one of those in which there's even a matter of fact about whether or not any sentient experimenters exist at all. And of course in the event that there *isn't* any matter of fact about

[7] Of course, the hypothesis that the equations of motion are always exactly right is *also* the sort of thing that Daneri, Loinger, and Prosperi and all those other guys in note 16 of Chapter 5 took themselves to be adherents of.

The trouble is that (astonishingly) it never seems to have *occurred* to *those* guys that it follows from that hypothesis that experiments almost never have outcomes; and so none of them ever worried about how to come to terms with that; and so none of them ever entered into the sorts of considerations that we're in the midst of here.

[8] This very nice way of putting the problem is due to Joshua Newman.

whether or not any sentient experimenters exist, then it becomes unintelligible even to inquire (as we've been doing here) about what sorts of things such experimenters will *report.*

And then (as far as I can tell) all bets are off.

And so it seems to me *not* to be entertainable, in any of the ways we've talked about so far, or in any *other* way I know of, that the linear quantum-mechanical equations of motion are the true and complete equations of motion of the whole world. And that's that.

But there are nonetheless interesting things to be learned about the measurement problem (things that it will be well to bear in mind in connection with the problems we ran into at the end of the last chapter, and in connection with problems we *will* run into at the end of the *next* chapter) in this stuff about what superpositions feel like.

What that stuff shows, I think, is that precisely that feature of those equations which makes it *clear* that they cannot possibly be the true and complete equations of motion of the whole world (that is: their *linearity*) *also* makes it radically *un*clear *how much* of the world and *which parts* of the world those equations possibly *can* be the true and complete equations of motion of.

What I think it shows (to put it another way) is that there can be *no such thing* as a definitive list of what there have absolutely got to be matters of fact about which is scientifically fit to serve as an "observational basis" from which all attempts at fixing quantum mechanics up must *start out.*

What I think it shows is that what there are and what there aren't determinate matters of fact about, even in connection with the most mundane and everyday macroscopic features of the external physical world, and even in connection with the most mundane and everyday features of *our own mental lives,* is something which we shall ultimately have to *learn* (in some part) from whatever turns out to be the *best* way of fixing quantum mechanics up.[9]

[9] But note that that learning will be no straightforward matter, since one of the things that all this raises difficult questions about is the very business of *seeking out* the best way of fixing quantum mechanics up!

The Dynamics Almost by Itself

Let's start again.

Suppose that there's just one world. And suppose that there's just one complete story of the world that's true.

And suppose that quantum-mechanical state vectors are complete descriptions of physical systems. And suppose that the dynamical equations of motion are always exactly right.

And suppose that we should like to insist, as a matter of principle, that healthy people can correctly report whether or not they themselves have any determinate belief about (say) the position of some particular pointer.

Then (since the dynamical equations of motion entail that healthy people in superpositions of brain states corresponding to different beliefs about the position of some particular pointer will with certainty report that *they have some determinate belief* about the position of that pointer) there's going to have to be something funny about how mental states supervene on brain states.[10]

Let's see if we can cook up something funny like that.

Think of *h* when she's about to measure the color of the hard electron, when she's in her "ready" state. When the measurement is over, the physical state of *h* and her measuring device and the electron is going to be the one in (6.1). That's what's dictated, with certainty, by the deterministic equations of motion.

Suppose, however, that all that's true, but that the evolution of *h*'s *mental* state in the course of a measurement like this one is explicitly *probablistic.* Here's how things would go in this particular case: *h* starts out (with certainty) in the mental state associated with $|\text{ready}\rangle_h$, and she ends up (with equal probabilities) either in the mental state associated with $|\text{believes } e \text{ black}\rangle_h$ or in the mental state associated with $|\text{believes } e \text{ white}\rangle_h$.[11] What's *certain* about how she ends up, though, is that she ends up (just as she testifies she

[10] That is, it's going to have to be the case that somebody's believing that such-and-such is *not* identical with some particular state of that person's *brain* (the state we've been calling $|\text{believes such-and-such}\rangle$) obtaining.

[11] It's obvious how this ought to be generalized: In the event that the initial state of the electron is $a|\text{white}\rangle + b|\text{black}\rangle$, and in the event that *h* measures that electron's color, then (as above) *h* will start out, with certainty, in the mental state

does) with some perfectly determinate belief about what the color of the electron is.[12]

So far so good. Let's try to take it a little further.

Whatever belief *h* *does* end up with, when (6.1) obtains, is necessarily going to be a false belief. But there are very natural ways of cooking things up so as to guarantee that that belief will nonetheless have an important kind of *effective validity*, at least in so far as *h* is concerned;[13] there are ways of cooking things up (that is) so as to guarantee that the future evolution of *h*'s mental state will proceed, in general, exactly *as if h*'s beliefs *were* true.

Here's what I mean.

Suppose that the mental state that *h* ends up in when (6.1) obtains (call that time *t*) happens to be the one associated with $|$believes *e* black$\rangle_h$, and suppose that she subsequently repeats that color measurement (with another color measuring device) on that same electron. When that's done, the physical state of things is going to be the one in (6.2), and *h* will with certainty (on this proposal) end up in the mental state associated with $|$believes outcome of first measurement is "black" and believes outcome of second measurement is "black"$\rangle_h$.[14] And that's precisely how *h*'s mental state *would* have ended up, with certainty (on this proposal), in the event that her belief that the electron was *black* at time *t* (which was false) had been *true*.

And suppose (just as above) that a state like the one in (6.1)

associated with $|$ready$\rangle_h$, and she'll end up in the mental state associated with $|$believes *e* white$\rangle_h$ with probability $|a|^2$, and she'll end up in the mental state associated with $|$believes *e* black$\rangle_h$ with probability $|b|^2$.

[12] This sort of thing was first suggested quite a long time ago, but for somewhat different reasons (for reasons which had nothing to do with what the equations of motion dictate about what it feels like to be in a superposition) by Bernard d'Espagnat (1971).

[13] Of course, there was a sense in which it seemed right to say that *h effectively knows* what the color of the electron is, when (6.1) obtains, on the *bare* theory too. But what we're talking about now will amount to something a good deal stronger than that.

[14] And of course in the event that *h*'s mental state at *t* happens to be the one associated with $|$believes *e* white$\rangle_h$, then *h* will with certainty end up in the mental state associated with $|$believes outcome of first measurement is "white" and believes outcome of second measurement is "white"$\rangle_h$.

obtains, and suppose that *h*'s mental state happens to be the one associated with |believes *e* black$\rangle_h$, and suppose that she subsequently carries out a measurement of the *hardness* of that same electron; then, when *that's* done, the physical state of things is going to be

.6) $1/\sqrt{4}$\{(|believes outcome of first measurement is "black" and believes outcome of second measurement is "hard"$\rangle_h$|"black"$\rangle_{m1}$|"hard"$\rangle_{m2}$|hard$\rangle_e$)

+ (|believes outcome of first measurement is "black" and believes outcome of second measurement is "soft"$\rangle_h$|"black"$\rangle_{m1}$|"soft"$\rangle_{m2}$|soft$\rangle_e$)

+ (|believes outcome of first measurement is "white" and believes outcome of second measurement is "hard"$\rangle_h$|"white"$\rangle_{m1}$|"hard"$\rangle_{m2}$|hard$\rangle_e$)

− (|believes outcome of first measurement is "white" and believes outcome of second measurement is "soft"$\rangle_h$|"white"$\rangle_{m1}$|"soft"$\rangle_{m2}$|soft$\rangle_e$)\}

(where *m*1 is the color measuring device and *m*2 is the hardness measuring device), and the probability that *h* will end up in the mental state associated with |believes outcome of first measurement is "black" and believes outcome of second one is "hard"$\rangle_h$ will be ½, and the probability that she will end up in the mental state associated with |believes outcome of first measurement is "black" and believes outcome of second measurement is "soft"$\rangle_h$ will be ½, and the probability of her ending up in the mental states associated with either |believes outcome of first measurement is "white" and believes outcome of second measurement is "hard"$\rangle_h$ or |believes outcome of first measurement is "white" and believes outcome of second measurement is "soft"$\rangle_h$ will be 0.

And suppose that a state like the one in (6.4) obtains and that *h*'s mental state happens to be the one associated with |believes 1 black$\rangle_h$, and suppose that *h* now carries out a measurement of the color of electron 2 (in which case the physical state of the world will become the one in (6.5)). Then (on this proposal) the probability of *h*'s ending up in the mental state associated with |believes 1 black and 2 black⟩$_h$ will be ½, the probability of *h*'s ending up in the mental state associated with |believes 1 black and 2 white⟩$_h$ will be ½, and the probabilities of *h*'s ending up in the mental states associated with either |believes 1 white and 2 black⟩$_h$ or |believes 1 white and 2 white⟩$_h$ will both be 0.

And so on.[15]

That (technically) will do the trick. On this proposal, quantum-mechanical wave functions are complete descriptions of the physical states of things, and those wave functions invariably evolve in perfect accordance with the dynamical equations of motion, and it makes no physical difference at all what *basis* we choose to write those wave functions *down* in,[16] and measurements carried out by sentient observers (that is: by observers with *minds*) invariably have determinate *outcomes* in the minds of those observers, and the statistical *distributions* of those outcomes will be the usual quantum-mechanical ones, and there isn't anything mysterious about how *probabilities* come up in this theory,[17] and the *reports* of sentient observers about their own mental states will invariably, on this proposal, be *correct*.

[15] What's been said so far (a slightly more detailed account of which, by the way, can be found in Albert and Loewer, 1988) doesn't amount to a completely general set of laws of the evolution of mental states; but laws like that can be cooked up, and they can be cooked up in such a way as to guarantee that everything I've said about them so far will be true.

[16] Of course, there will (on this picture, and on every way of attempting to make sense of quantum mechanics) be some particular basis of brain states which correspond to (as it were) "eigenstates of mentality"; but what basis that *is* will by no means be a matter of conventional choice; what basis that is will entirely depend (rather) on the *physical structure* of the *brains* in question. The brain state that corresponds to believing that a certain electron is black, for example, will presumably be the one which (purely in virtue of the dynamical equations of motion) disposes its owner to respond to an utterance like "What color do you believe the electron to be?" with an utterance like "I believe the electron to be black." And of course what brain state *that* will be will be a completely basis-independent, straightforwardly physical question!

[17] The way that probabilities come up in *this* theory, after all, is that they get *put* into it by *fiat;* and that fiat stipulates that those probabilities are to be understood in precisely the conventional way.

And of course this view of the world is a thoroughly realist one (that is: this view entails that there is invariably a single correct objective description of the entire physical and mental universe, even if nobody happens to know what that description *is*); and this view (even though it's an explicitly dualist view) entails that the mental parts of the world have no effects whatever on its physical parts (that is: this view isn't at all like any of the dualist theories of *collapse, this* view entails that the physical world is *causally closed*).

But the dualism of this sort of a picture is nonetheless pretty bad. On this proposal (for example) all but one of the terms in a superposition like the one in (6.1) represent (as it were) *mindless hulks;* and *which one* of those terms is *not* a mindless hulk can't be deduced from the physical state of the world, or from the outcome of any sort of an experiment; and it will follow from this proposal that most of the people we take ourselves to have met in our lives have as a matter of fact *been* such hulks, and not really people (not really animate, that is) at all!

Here's a way to partly fix *that* up:

Suppose that every sentient physical system there is is associated not with a single mind but rather with a *continuous infinity* of minds; and suppose (this is part of the proposal too) that the measure of the infinite subset of those minds which happen to be in some particular mental state at any particular time is equal to the square of the absolute value of the coefficient of the brain state associated with that mental state, in the wave function of the world, at that particular time (so that, for example, when states like (6.1) obtain, half of *h*'s continuous infinity of minds will believe that the electron is black, and half of them will believe that the electron is white).

The time evolution of each individual mind, on this proposal, is precisely the probabilistic one described above (the one that we cooked up for the single-mind proposal), but since (on *this* proposal) there are always a continuous infinity of minds (or else no minds at all) in any particular mental state, the evolution of the minds of any particular sentient observer *as a set* is invariably (that is: with probability 1) going to be *deterministic.* Moreover, at any particular instant, the mental states of the minds of any particular observer will necessarily be distributed in accordance with the prescription of the last paragraph. So this proposal is going to entail that what you might call the "global" mental state of every sentient being *is* uniquely fixed by the physical state of the world.[18]

And there's something else about this kind of a picture that's nice: this kind of a picture is *local.* That's surprising. That's precisely the sort of thing that Bell's theorem was thought to have ruled out. Let's see how it works.

Consider an EPR-type state:

$$7) \qquad |\text{black}\rangle_1|\text{white}\rangle_2 - |\text{white}\rangle_1|\text{black}\rangle_2$$

and suppose that electron 1 is located at point 1 and that electron 2 is located at point 2 and that an observer named *h*1 (located at point 1) measures some spin observable of electron 1 and that an observer named *h*2 (located at point 2) measures some spin observable (not necessarily the same one) of electron 2.

What Bell proved is that there can't be any local way of accounting for the observed correlations between the outcomes of measurements like that; but of course (and this is the crux of the whole business) the idea that there ever *are* matters of fact about the "outcomes" of a pair of measurements like that is just what *this* sort of a picture *denies!*

Let's go through it carefully.

At the conclusion of a pair of measurements like the one just described, on *this* picture, the state of the world is going to be a superposition of states, in each of which each of those two measurements have one or the other of their two different possible outcomes. And at that point, on this picture, no matter what spin observable of electron 1 gets measured by *h*1 and no matter what spin observable of electron 2 gets measured by *h*2, half of *h*1's minds are going to believe that the outcome of whatever measurement *she* did was +1, and the other half of her minds are going to

[18] This is the so-called many-minds interpretation of quantum mechanics, which was first proposed by Barry Loewer and myself (Albert and Loewer, 1988).

believe that the outcome of whatever measurement she did was -1, and half of *h*2's minds are going to believe that the outcome of whatever measurement *she* did was $+1$, and the other half of her minds are going to believe that the outcome of whatever measurement she did was -1. (The reader will have no trouble in explicitly confirming all this, and in confirming that none of this depends on the *time order* in which the two measurements get carried out.)

And this is where things get a little more subtle.

What it's hard not to do, at first, at this point in the story, is to imagine that there are matters of fact (when this sort of a superposition obtains) about the degree to which the states of the minds of *h*1 and the states of the minds of *h*2 are *correlated* with one another.

But the thing is that there *aren't* any matters of fact about anything like that.

All that ever actually *happens* (insofar as any question of *correlations* is concerned) is that at the point (later on) when *h*1 actually *communicates* with *h*2 (and that communication is of course going to be mediated by some *local* interaction and governed by the *local* equations of motion), then each one of each of these two observers' minds will develop some particular *belief* about whether or not the outcome of the *other* observer's measurement was correlated or *anti*-correlated with the outcome of *her own*. And the *probability* of developing any particular such belief (for each of the two observers separately) is going to be precisely the usual quantum-mechanical one. And (as I said before) there simply isn't going to *be* any matter of fact about whether or not the outcomes of these two measurements, or the beliefs of these two observers, are ever "really" correlated with one another.

An Epistemological Remark

One of the things that the many-minds interpretation entails (as I've already mentioned) is that the beliefs of any sentient observer about the overall quantum state of the world will typically be mistaken.

Nothing, even in principle, can be done about that. No matter how much the observer in question knows of what the true laws of the world are, and no matter what observables she is capable of measuring, there can't be any experimental means whatever (as a little reflection will show) of reliably finding out what the overall quantum state of the world is, or (for that matter) what the quantum state of anything *in* the world is.

The sum total of what any such observer can conclude about the overall quantum state of the world (or, more precisely: the sum total of what any particular one of such an observer's *minds* can conclude about the overall quantum state of the world), from the outcomes of whatever experiments she does, is that that state (whatever it is) is not orthogonal (that is: not *perfectly* orthogonal) to the *effective* state that those outcomes *pick out.* And that's all.

And that's not much.

And that fact has curious consequences. It turns out, for example, that the Lorentz-covariance of the dynamical equations of motion of relativistic quantum field theories requires that the state that's associated with the *vacuum* in theories like that is necessarily not quite perfectly orthogonal to states in which there are electrons and baseballs and people and buildings (and all the other stuff we're used to) around.

And so what we've just been talking about is going to entail that on relativistic-field-theoretic versions of (say) a many-minds picture, nothing in our empirical experience (that is: nothing about the histories of our phenomenal states) is incompatible with the hypothesis that the quantum state of the universe is (for now and for all time) that *vacuum* state!

And that will throw an odd light (for example) on questions about where the universe initially *came* from.

But going through the details of all this would require a more technical discussion than I want to get into just now.[19]

[19] I hate lines like that. But let the reader take note that this is the only one of them in this book.

Anyway, a slightly less incomplete account of this stuff (together with some further references and some remarks about how these ideas are and aren't related to the literature on the possibility that the universe is a vacuum fluctuation) can be found in a little paper of mine from a couple of years ago (Albert, 1988).