# ••• 4 •••

# The Measurement Problem

Now I want to begin to worry in earnest about whether or not the dynamics says the same thing as the postulate of collapse says about what happens to the state vector of a physical system when the system gets measured. Here's what looked worrisome back in Chapter 2: the dynamics (which is supposed to be about how the state vectors of physical systems evolve *in general*) is fully deterministic, but the collapse postulate (which is supposed to be about how the state vector of a system evolves when it comes in contact with a measuring device) isn't; and so it isn't clear precisely how the two can be consistent.

Let's figure out what the dynamics says about what happens when things get measured.

Suppose that everything in the world always evolves in accordance with the dynamical equations of motion. And suppose that we have a device (which operates in accordance with those equations, just like everything else does) for measuring the hardness of an electron; and suppose that that device works like this: The device has a dial on the front, with a pointer; and the pointer has three possible positions. In the first position the pointer points to the word "ready," and in the second position it points to the word "hard," and in the third it points to the word "soft." Electrons are fed into one side of the device and come out the other, and in the course of passing through (if the device is set up right, with the pointer initially in its "ready" position) they get their hardnesses measured, and the outcomes of those measurements get recorded in the final position of the pointer (figure 4.1).

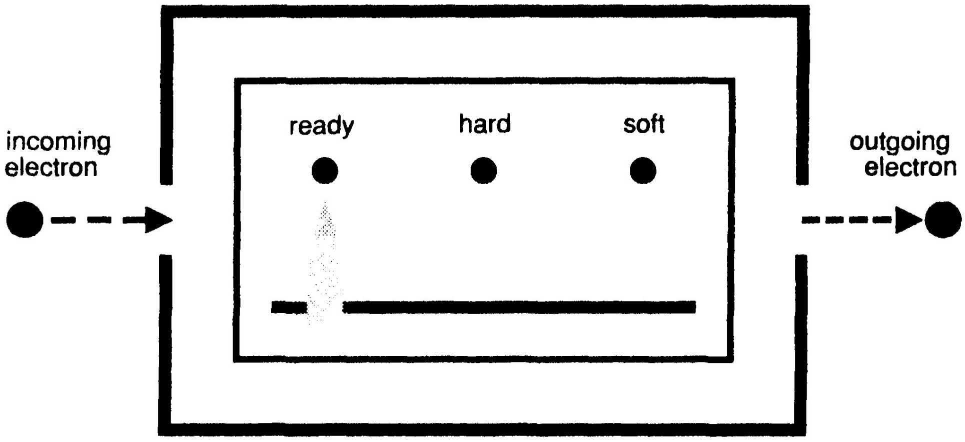

*Figure 4.1*

If the device is set up right, and if the dynamics is always true, then (to put all this another way) the dynamical equations of motion entail that it behaves like this:

1) $|\text{ready}\rangle_m|\text{hard}\rangle_e \rightarrow |\text{“hard”}\rangle_m|\text{hard}\rangle_e$

and

2) $|\text{ready}\rangle_m|\text{soft}\rangle_e \rightarrow |\text{“soft”}\rangle_m|\text{soft}\rangle_e$

That is: if the device (whose state vector is labeled with subscript *m*) is initially in the ready state, and if an electron (whose state vector is labeled with subscript *e*) that is hard gets fed through it, then the device ends up in the state wherein the pointer is pointing at “hard”; and if the device is initially in the ready state, and if a *soft* electron gets fed through it, then the device ends up in the state wherein the pointer is pointing at “soft.” That's what it *means* for a measuring device for hardness to be a good one and to be set up right.

Now (still supposing that the dynamics is always true), consider what happens if this device (the one for measuring *hardness*) is set up right, and is in its ready state, and a *black* electron is fed into it. It turns out that (4.1) and (4.2), and the fact that the dynamical equations of motion are invariably linear, suffice by themselves to figure that out: The initial state of the electron and the measuring device is

.3) $$|\text{ready}\rangle_m|\text{black}\rangle_e = |\text{ready}\rangle_m\{\tfrac{1}{\sqrt{2}}|\text{hard}\rangle_e + \tfrac{1}{\sqrt{2}}|\text{soft}\rangle_e\} \\ = \tfrac{1}{\sqrt{2}}|\text{ready}\rangle_m|\text{hard}\rangle_e + \tfrac{1}{\sqrt{2}}|\text{ready}\rangle_m|\text{soft}\rangle_e$$

which is precisely $\frac{1}{\sqrt{2}}$ times the initial state in (4.1) plus $\frac{1}{\sqrt{2}}$ times the initial state in (4.2). So, since (by hypothesis) the dynamical equations of motion entail that $|\text{ready}\rangle_m|\text{hard}\rangle_e$ evolves as in (4.1) and that $|\text{ready}\rangle_m|\text{soft}\rangle_e$ evolves as in (4.2) (that is: since this device is set up right, and since it's a good measuring device for hardness), it follows from the linearity of those equations that the state in (4.3), when the measuring device gets switched on, will necessarily evolve into

.4) $$\tfrac{1}{\sqrt{2}}|\text{“hard”}\rangle_m|\text{hard}\rangle_e + \tfrac{1}{\sqrt{2}}|\text{“soft”}\rangle_m|\text{soft}\rangle_e$$

That's how things end up, with certainty, according to the dynamics.[1]

And the way things end up according to the postulate of *collapse* (when you start with (4.3)) is

5) *either* $|\text{“hard”}\rangle_m|\text{hard}\rangle_e$ (with probability ½)
*or* $|\text{“soft”}\rangle_m|\text{soft}\rangle_e$ (with probability ½)

[1] This result often strikes people as mysteriously easy. The intuition is that measuring devices for hardness like the one described here must be extremely complicated contraptions (especially if you look at them on the level, say, of their constituent atoms) and must have extremely complicated equations of motion, the solution of which must be an extremely complicated matter. All of that is true. What simplifies things here is the fact that however complicated those equations may be, (4.1) and (4.2) are surely solutions of them (since the contraption we're dealing with is, by hypothesis, and whatever *else* it may be, a good measuring device for the hardness of an electron), and they (the equations) are surely *linear;* and those two facts are enough by themselves to insure that if a black electron is fed into the device, then those equations will entail that things will end up in the state in (4.4).

And the trouble is that (4.4) and (4.5) are measurably different situations.[2]

The state described in (4.5) is the one that's right; it is (as a matter of empirical fact) how things *do* end up when you start with (4.3).

The state described in (4.4) is *not* how things end up;[3] (4.4) is something very strange. It's a superposition of one state in which the pointer is pointing at "hard" and another state in which the pointer is pointing at "soft"; it's a state in which (on the standard way of thinking) *there is no matter of fact* about where the pointer is pointing.[4]

Let's make this somewhat sharper. Suppose that a human observer enters the picture, and *looks* at the measuring instrument (when the measurement is all done) and *sees* where the pointer is pointing. Let's figure out what the dynamics will say about that.

Suppose, then (just as we did before), that literally every physical system in the world (and this now includes human beings; and it

[2] Perhaps this ought to be expanded on a bit. The point is that there are (in accordance with the postulates of quantum mechanics that were laid out in Chapter 2) necessarily measurable properties of the state in (4.4) whereby it can, in principle, be experimentally distinguished from either of the states in (4.5) and, as a matter of fact, from any other state whatever. There will be a good deal to say, later on, about precisely *what* those properties *are* (they're complicated ones, and their measurement will in general be extremely difficult); what's important for the moment is simply that those properties exist.

[3] It isn't, at any rate, according to the conventional wisdom about these matters; but there will be much more to say about this later on.

[4] Something ought to be mentioned in passing here, something that will turn out to be important later on.

What we've just discovered is that there is a certain fundamental effect of the carrying out of a measurement (namely: the emergence of some definite *outcome* of the measurement; the emergence of some *matter of fact* about precisely *what* the outcome of the measurement *is*) which is not predicted by the dynamical equations of motion.

But consider another effect of the carrying out of a measurement, one which we first described in the course of our discussions of hardness and color in Chapter 1: The carrying out of a measurement is disruptive of the values of observables of the measured system which are incompatible with the observable that gets measured. It turns out that the dynamical equations of motion *do* predict *that*.

includes the brains of human beings) always evolves in accordance with the dynamical equations of motion; and suppose that a black electron is fed through a measuring device for hardness that's set up right and that starts out in its ready state (so that the state of the electron and the device is now the one in (4.4)); and suppose that somebody named Martha comes along and looks at the device; and suppose that Martha is a competent observer of the positions of pointers.

Being a "competent observer" is something like being a measuring device that's set up right: What it means for Martha to be a competent observer of the position of a pointer is that whenever Martha looks at a pointer that's pointing to "hard," she eventually comes to *believe* that the pointer is pointing to "hard"; and that whenever Martha looks at a pointer that's pointing to "soft," she eventually comes to believe that the pointer is pointing to "soft" (and so on, in whatever direction the pointer may be pointing). What it means (to put it somewhat more precisely) is that the dynamical equations of motion entail that Martha (who is a physical system, subject to the physical laws) behaves like this:

6) $$|\text{ready}\rangle_o|\text{ready}\rangle_m \rightarrow |\text{``ready''}\rangle_o|\text{ready}\rangle_m$$

and

$$|\text{ready}\rangle_o|\text{``hard''}\rangle_m \rightarrow |\text{``hard''}\rangle_o|\text{``hard''}\rangle_m$$

and

$$|\text{ready}\rangle_o|\text{``soft''}\rangle_m \rightarrow |\text{``soft''}\rangle_o|\text{``soft''}\rangle_m$$

In these expressions, $|\text{ready}\rangle_o$ is that physical state of Martha's brain in which she is alert and in which she is intent on looking at the

Look, for example, at the evolution from (4.3) to (4.4). Equation (4.3) is an eigenstate of the *color* of the electron whose hardness is about to be measured; but (4.4) (which is the state following the interaction of the electron with a good measuring device for the hardness, according to the equations of motion) *isn't.* If (4.4) is written out in terms of eigenstates of the color of the electron (which the reader can now easily do), *it* turns out to be a superposition of two states (with equal coefficients), in one of which the electron is black and in the *other* of which the electron is white.

pointer and finding out where it's pointing; $|\text{“ready”}\rangle_o$ is that physical state of Martha's brain in which she believes that the pointer is pointing to the word “ready” on the dial; $|\text{“hard”}\rangle_o$ is that physical state of Martha's brain in which she believes that the pointer is pointing to the word “hard” on the dial; and $|\text{“soft”}\rangle_o$ is that physical state of Martha's brain in which she believes that the pointer is pointing to the word “soft” on the dial.[5]

Let's get back to the story. The state of the electron and the measuring device (at the point where we left off) is the strange one in (4.4). And now in comes Martha, and Martha is a competent observer of the position of the pointer, and Martha is in her ready state, and Martha looks at the device. It follows from the linearity of the dynamical equations of motion (if those equations are right), and from what it means to be a competent observer of the position of the pointer, that the state when Martha's done is with certainty going to be

7) $$1/\sqrt{2}|\text{“hard”}\rangle_o|\text{“hard”}\rangle_m|\text{hard}\rangle_e + 1/\sqrt{2}|\text{“soft”}\rangle_o|\text{“soft”}\rangle_m|\text{soft}\rangle_e$$

That's what the dynamics entails.

And of course what the postulate of *collapse* entails is that when Martha's all done, then

8) *either* $|\text{“hard”}\rangle_o|\text{“hard”}\rangle_m|\text{hard}\rangle_e$ (with probability ½)
*or* $|\text{“soft”}\rangle_o|\text{“soft”}\rangle_m|\text{soft}\rangle_e$ (with probability ½)

is going to obtain.

And (4.7) and (4.8) are empirically different. The state described in (4.8) is the one that's right; (4.7) is unspeakably strange. The state described in (4.7) is at odds with what we know of ourselves

[5] It hardly needs saying that this is an absurdly oversimplified description of Martha's brain, and that this is an absurdly oversimplified account of the ways in which mental states are generally supposed to supervene on brain states; but all that turns out not to make any difference (not at *this* stage of the game, anyway). We can fill in the details whenever we want, to whatever extent we want. They won't change the arguments.

by *direct introspection.* It's a superposition of one state in which Martha thinks that the pointer is pointing to "hard" and another state in which Martha thinks that the pointer is pointing to "soft"; *it's a state in which there is no matter of fact about whether or not Martha thinks the pointer is pointing in any particular direction.*[6]

And so things are turning out badly. The dynamics and the postulate of collapse are flatly in contradiction with one another (just as we had feared they might be); and the postulate of collapse seems to be right about what happens when we make measurements, and the dynamics seems to be bizarrely *wrong* about what happens when we make measurements; and yet the dynamics seems to be *right* about what happens whenever we *aren't* making measurements; and so the whole thing is very confusing; and the problem of what to do about all this has come to be called "the problem of measurement."

We shall be thinking about that for the rest of this book.

[6] This isn't anything like a state in which Martha is, say, *confused* about where the pointer is pointing. *This* (it deserves to be repeated) is something *really strange.* This is a state wherein (in the language we used in Chapter 1) it isn't right to say that Martha believes that the pointer is pointing to "hard," and it isn't right to say that Martha believes that the pointer is pointing to "soft," and it isn't right to say that she has *both* of those beliefs (whatever *that* might mean), and it isn't right to say that she has neither of those beliefs.

# ... 5 ...

# The Collapse of the Wave Function

## The Idea of the Collapse

The measurement problem was first put in its sharpest possible form in the 1930s, by John von Neumann, in an extraordinary book called *Mathematical Foundations of Quantum Mechanics* (von Neumann, 1955). It looked to von Neumann as though the only thing that could possibly be done about the measurement problem was to bite the bullet, and admit that the dynamics is simply wrong about what happens when measurements occur, and nonetheless right about everything else. And so what he concluded was that there must be two fundamental laws about how the states of quantum-mechanical systems evolve:

I. When no measurements are going on, the states of all physical systems invariably evolve in accordance with the dynamical equations of motion.

II. When there *are* measurements going on, the states of the measured systems evolve in accordance with the postulate of collapse, *not* in accordance with the dynamical equations of motion.

But this clearly won't do. Here's the trouble: What these laws actually *amount to* (that is: what they actually *say*) will depend on the precise meaning of the word *measurement* (because these two laws entail that *which one* of them is being *obeyed* at any given moment depends on whether or not a "measurement" is being *carried out* at that moment). And it happens that the word *mea-*

*surement* simply doesn't have any absolutely precise meaning in ordinary language; and it happens (moreover) that von Neumann didn't make any attempt to cook up a meaning for it, either.

And so those laws, as von Neumann wrote them down, simply don't determine exactly how the world behaves (which is to say: they don't really amount to prospective fundamental "laws" at all).

And there has consequently been a long tradition of attempts to figure out how to write them down in such a way that they *do*.

Here's where things stand: Suppose that a certain system is initially in an eigenstate of observable *A*, and that a measurement of observable *B* is carried out on that system, and that *A* and *B* are incompatible with one another. What we know with absolute certainty, by pure introspection, is that by the time that measurement is all done, and a sentient observer has looked at the measuring device and formed a conscious impression of how that device presently appears and what it presently indicates, then some wave function must already have violated the dynamical equations of motion and collapsed. What we need to do is to figure out precisely when that collapse occurs.

Let's try to guess.

Perhaps the collapse always occurs precisely at the last possible moment; perhaps (that is) it always occurs precisely at the level of consciousness,[1] and perhaps, moreover, consciousness is always the agent that brings it about.

Put off the temptation to dismiss this as nonsense just for long enough to see what it amounts to.

On this proposal (which is due to Wigner, 1961), the correct laws of the evolution of the states of physical systems look something like this: All physical objects almost always evolve in strict accordance with the dynamical equations of motion. But every now and then, in the course of some such dynamical evolutions (in the course

[1] This isn't a way of saying that there's anything *illusory* about the collapse; it's just a way of saying at precisely what point the collapse (which is a *physical* process) occurs, a way of saying precisely what sorts of processes precipitate collapses.

of *measurements*, for example), the brain of a sentient being may enter a state wherein (as we've seen) states connected with various different *conscious experiences* are superposed; and at such moments, the *mind* connected with that brain (as it were) *opens its inner eye*, and *gazes* on that brain, and that causes the entire system (brain, measuring instrument, measured system, everything) to collapse, with the usual quantum-mechanical probabilities, onto one or another of those states;[2] and then the eye closes, and everything proceeds again in accordance with the dynamical equations of motion until the next such superposition arises, and then that mind's eye opens up again, and so on.

This proposal entails that there are two fundamentally different sorts of physical systems in the world:

A. *Purely physical* systems (that is: systems which *don't* contain sentient observers). These systems, so long as they remain isolated from outside influences, always evolve in accordance with the dynamical equations of motion.

B. *Conscious* systems (that is: systems which *do* contain sentient observers). These systems evolve in accordance with the more complicated rules described above.[3]

But the trouble here is pretty obvious too: How the physical state of a certain system evolves (on this proposal) depends on whether or not that system is *conscious*; and so in order to know precisely how things physically behave, we need to know precisely what is

[2] Here's an example: An observer carries out a measurement of the hardness of a black electron. Eventually (when the measuring device has done its work, and the observer has looked at the device), things get to be in the state in equation (4.7), and then the mind's eye of the observer opens, gazes upon her brain, and causes a collapse, with equal probabilities, onto either the first or the second of the terms in that state.

[3] This (needless to say) amounts to a full-blown radically interactive mind-body dualism. Wigner thought that this sort of dualism turns out (ironically) to be a *necessary consequence of physics:* he thought that there was a physical job to be done in the world (the job of collapsing wave functions) which could only be done by a not-purely-physical thing.

conscious and what isn't. What this "theory" predicts (that is: what "theory" it *is*) will hinge on the precise meaning of the word *conscious;* and that word simply doesn't have any absolutely precise meaning in ordinary language; and Wigner didn't make any attempt to make up a meaning for it; and so all this doesn't end up amounting to a genuine physical theory either.

Let's try another guess. Perhaps the collapse occurs at the level of *macroscopicness* (not at the last possible moment but, as it were, at the last reasonable moment). On this proposal (which has lots of originators and lots of adherents), the correct laws of the evolution of the states of physical systems look something like this: All physical objects almost always evolve in strict accordance with the dynamical equations of motion. But every now and then, in the course of some such dynamical evolutions (in the course of measurements, for example), it comes to pass (as we've seen) that two macroscopically different conditions of a certain system (two different orientations of a pointer, say) get superposed, and at that point, as a matter of fundamental physical law, the state of the entire system collapses, with the usual quantum-mechanical probabilities, onto one or another of those macroscopically different states.[4] Then everything proceeds again in accordance with the dynamical equations of motion until the next such superposition arises, and then another such collapse takes place, and so on.

There are two sorts of physical systems in the world according to this proposal too:

A′. Purely *microscopic* systems (that is: systems which *don't* contain macroscopic subsystems). These systems, so long as they remain isolated from outside influences, always evolve in accordance with the dynamical equations of motion.

B′. *Macroscopic* systems (that is: systems which *do* contain macroscopic subsystems). These systems evolve in accordance with the more complicated rules described above.

[4] Here's an example: Somebody carries out a measurement of the hardness of a black electron. Eventually (when the measuring device has done its work), things get to be in the state in equation (4.4), and then that state collapses, with equal probabilities, onto either the first or the second of the terms in that state.

The trouble *here* is going to be about the meaning of *macroscopic.* And that will put us back (for the third time now) where we started.[5] And so all this is getting us nowhere.

We need to find a less ambiguous way to talk.

## A Digression on Why It's Hard to Settle the Question Empirically

Let's see if we can settle the question empirically.

Here's the idea: The collapse of the wave function (whatever else it might turn out to be) is, after all, a physical event, with physical consequences; and those consequences must in principle be detectable; and so the question of precisely where and when collapses occur (which is what we've been merely guessing about over the last few pages) must in principle be answerable, with certainty, by means of the right sorts of experiments.

Suppose (for example) that I have a theory about the collapse; and suppose that my theory entails that if I pass a black electron through a measuring device for hardness like the one described in Chapter 4, then at precisely the moment when the state of the electron and the measuring device becomes (in accordance with the dynamics) the one in equation (4.4), that state collapses, with the usual quantum-mechanical probabilities, onto one or the other of the two terms in that state. And suppose that my friend has another theory about the collapse; and suppose that *her* theory entails that the collapse doesn't happen until some particular *later* moment, some moment *farther on* in the measuring process (only when, say, a human retina gets involved, or an optic nerve, or a brain, or whatever). And suppose that we should like to test these two theories against one another by means of an experiment.

Here's how to start: Feed a black electron into a measuring device for hardness and give it enough time to pass all the way through. If *my* theory is right, then the state of the electron and the measur-

[5] There is, as a matter of fact, an astonishingly long and bombastical tradition in theoretical physics of formulating these sorts of guesses about precisely when the collapse occurs in language which is so imprecise as to be (as we've just seen) absolutely useless. Some of the words that come up in these guesses (besides *measurement* and *consciousness* and *macroscopic*) are *irreversible, recording, information, meaning, subject, object,* and so on.

ing device at that moment (the moment at which the electron has had just enough time to pass all the way through the device) ought to be

1) *either* $|\text{“hard”}\rangle_m|\text{hard}\rangle_e$ (with probability ½)
*or* $|\text{“soft”}\rangle_m|\text{soft}\rangle_e$ (with probability ½)

But if my *friend's* theory is right, *then* the state at that same moment ought to be

2) $1/\sqrt{2}|\text{“hard”}\rangle_m|\text{hard}\rangle_e + 1/\sqrt{2}|\text{“soft”}\rangle_m|\text{soft}\rangle_e$

just as the dynamics predicts (the *violation* of the dynamics, in my friend's theory, doesn't happen until later). And so now all we need to do is to figure out a way to distinguish, by means of a measurement, between the circumstances in (5.1) (wherein the pointer is pointing in some particular, but as yet unknown, direction) and the circumstances in (5.2) (wherein the pointer *isn't* pointing in any particular direction *at all*).

Let's figure out what sort of an observable we would need to measure in order to do that.

What if we were to measure the position of the tip of the pointer (that is: what if we were to measure *where* the pointer is *pointing*)?[6] Here's why that won't work: If my theory is right and, consequently, (5.1) obtains, then of course a measurement of the position of the tip of the pointer will have a fifty-fifty chance of finding the pointer in the "pointing-at-hard" state, and it will have a fifty-fifty chance of finding the pointer in the "pointing-at-soft" state (since, on my theory, the pointer is now already *in* one of those two states, each with probability ½). But if my *friend's* theory is right and, consequently, (5.2) obtains, *then* a measurement of the position of

[6] And note that we shall be supposing in what follows that *this* measurement (the measurement of the position of the tip of the pointer on the hardness measuring device) gets carried through all the way to the end: all the way to the point where some observer becomes aware of the outcome of that measurement. By the time all that gets done, even my friend's theory about the collapse (that is: *any* theory about the collapse whatever) will have to entail that a collapse has already occurred.

the tip of the pointer will have a fifty-fifty chance of *collapsing* the state vector of the pointer onto the "pointing-at-hard" state, and a fifty-fifty chance of collapsing it onto the "pointing-at-soft" state.[7] And so the probability of any given outcome of a measurement of the position of the pointer will be the *same* on these two theories; and so this isn't the sort of measurement we're looking for.

What about measuring the hardness of the electron? That won't work either. If my theory is right and (5.1) obtains, then a measurement of the hardness of the electron will have a fifty-fifty chance of finding the electron in the hard state and a fifty-fifty chance of finding it in the soft state; and if my friend's theory is right and (5.2) obtains, then a measurement of the hardness of the electron will have a fifty-fifty chance of collapsing the state vector of the electron onto the hard state and a fifty-fifty chance of collapsing it onto the soft state. Again, the probabilities will be the same on both theories, and so this isn't the sort of measurement we're looking for, either.

What about measuring the *color* of the electron? That won't work either. On my theory, one of the two states in (5.1) now obtains and a measurement of the color of the electron in *either one* of those states will have a fifty-fifty chance of collapsing the state of the electron onto $|\text{black}\rangle_e$ and a fifty-fifty chance of collapsing the state of the electron onto $|\text{white}\rangle_e$. On my friend's theory, the state in (5.2) now obtains and a measurement of the color of the electron in *that* state will likewise have a fifty-fifty chance of collapsing the state of the electron onto $|\text{black}\rangle_e$ and a fifty-fifty chance of collapsing it onto $|\text{white}\rangle_e$.[8]

[7] Of course, in the event that the state of the pointer gets collapsed onto the "pointing-at-hard" state (onto $|\text{``hard''}\rangle_m$, that is), then the state of the electron will automatically get collapsed onto $|\text{hard}\rangle_e$; and in the event that the state of the pointer gets collapsed onto the "pointing-at-soft" state ($|\text{``soft''}\rangle_m$), then the state of the electron gets collapsed onto $|\text{soft}\rangle_e$. All of that follows from (5.2), together with the collapse postulate for composite systems, which was spelled out in Chapter 2.

[8] This is, as a matter of fact, something that we've noted before, in a slightly different language, in note 4 of Chapter 4. The reader can easily confirm it (as we mentioned there) by writing out the state in (5.2) in terms of eigenstates of the color of the electron.

Let's go back to the device. Consider an observable of the device which (for lack of any appropriate name) we can call zip. The eigenstates of zip are as follows:[9]

.3)
$$|\text{zip} = 0\rangle = |\text{ready}\rangle_m$$

$$|\text{zip} = +1\rangle = 1/\sqrt{2}|\text{“hard”}\rangle_m + 1/\sqrt{2}|\text{“soft”}\rangle_m$$

$$|\text{zip} = -1\rangle = 1/\sqrt{2}|\text{“hard”}\rangle_m - 1/\sqrt{2}|\text{“soft”}\rangle_m$$

Measuring zip won't work either, since (much like with the color of the electron) if either of the states in (5.1) obtains, or if the state in (5.2) obtains, a measurement of zip has a fifty-fifty chance of collapsing the state of the pointer onto the state $|\text{zip} = +1\rangle$ and a fifty-fifty chance of collapsing the state of the pointer onto the $|\text{zip} = -1\rangle$ state.[10]

[9] A few remarks are in order here.

First of all, it follows from property (5) of Hermitian operators (in Chapter 2) that some such observable as zip (that is: some such *Hermitian operator*, some operator with precisely these eigenstates) necessarily exists.

One of the eigenstates of zip (the zip = 0 state) is of course also an eigenstate of the pointer-position, but zip is nonetheless in general incompatible with the pointer-position: the zip = +1 state and the zip = −1 state are both superpositions of states in which the pointer is pointing in different directions.

The matrix for zip in the pointer-position basis

(that is: the basis in which $|\text{ready}\rangle_m = \begin{bmatrix}1\\0\\0\end{bmatrix}$, $|\text{“hard”}\rangle_m = \begin{bmatrix}0\\1\\0\end{bmatrix}$, and $|\text{“soft”}\rangle_m = \begin{bmatrix}0\\0\\1\end{bmatrix}$)

will be zip = $\begin{bmatrix}1 & 0 & 0\\0 & 0 & 1\\0 & 1 & 0\end{bmatrix}$

The reader may want to calculate the matrix for the pointer-position in this same basis (it's very easy), and to confirm that that latter matrix doesn't commute with the matrix for zip.

[10] Note, by the way, that everything that's been discovered here about the state in (5.2) could have been surmised, if we had been cleverer, right at the outset. The state in (5.2) is, after all, a *nonseparable* state of the electron and the measuring device, a state in which no observable of the electron alone or of the measuring device alone can *possibly* have any particular value.

Consider, finally, an observable of the composite system consisting of the measuring device *and* the electron, zip − color (that is: zip *minus* color). It turns out that the state in (5.2) *is* an eigenstate of zip − color (its associated eigenvalue is 0),[11] and that *neither* of the states in (5.1) are eigenvalues of zip − color. And so, if my theory is right and (5.1) obtains, then a measurement of zip − color (or, rather, a collection of such measurements, carried out on similarly prepared systems) will have a statistical distribution of various *different* possible outcomes; but if my *friend's* theory is right and (5.2) obtains, then the outcome of every measurement of zip − color will necessarily, with certainty, be 0. And so my friend's theory and my theory *can* be distinguished from one another, *empirically,* by means of measurements of zip − color.

And this, needless to say, is a very general sort of fact: any claim about precisely where and precisely when collapses occur can in principle be distinguished from any other one, empirically, by means of measurements more or less like this one.[12]

And so it would seem that the right way to *find out* precisely where and precisely when collapses occur must be just to go out and *perform* these sorts of measurements.

The trouble is that, for a number of reasons, that turns out to be an extraordinarily difficult business to actually carry through.

The nicest one of those reasons has to do with the ways that measuring devices, in virtue of their macroscopicness, necessarily interact with their *environments.*

Let's see how that works.

Suppose that a black electron is on its way into a hardness measuring device and that the measuring device is in its ready state. And suppose that (as in figure 5.1) there happens to be a *molecule of air* sitting just to the right of the pointer, so that if the pointer were to swing to its "hard" position then the molecule would get pushed to the middle of the dial, and if the pointer were to swing

[11] This can be confirmed by writing out the state in (5.2) in terms of eigenstates of color (for the electron) and of zip (for the pointer).

[12] This is the sort of measurable property referred to in note 2 of Chapter 4.

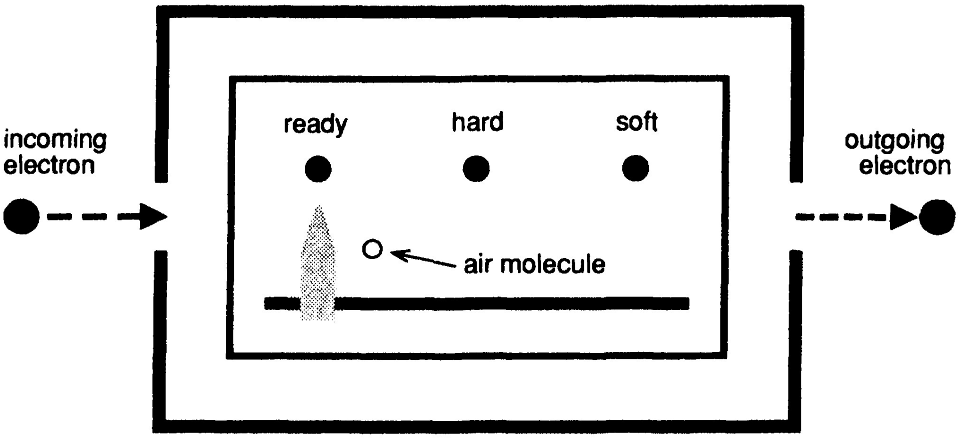

*Figure 5.1*

to its "soft" position then the molecule would get pushed to the right end of the dial.[13]

And now the electron passes through the hardness device, which swings the pointer, which pushes the air molecule (which behaves somewhat like a device for measuring the *position* of that pointer here);[14] and the reader (who is by now well versed in calculations about the behaviors of measuring devices) will have no trouble in confirming that the state of the electron and the hardness device and the molecule, when that's all done, *on my friend's theory,* is going to be

$$4)\qquad 1/\sqrt{2}(|\text{hard}\rangle_e|\text{“hard”}\rangle_m|\text{center}\rangle_a + |\text{soft}\rangle_e|\text{“soft”}\rangle_m|\text{right}\rangle_a)$$

where the vectors labeled with subscript *a* are state vectors of the air molecule.

That's how things stand at this point in the story, on my friend's theory, in the presence of an air molecule. Note that this state

[13] Note that we are attributing both a fairly well-defined velocity (rest) and a fairly well-defined position (near the pointer) to this molecule *at the same time* here. The mass of an air molecule is large enough that doing so need not conflict with the uncertainty relations.

[14] That is: the molecule is situated in such a way as to bring about a *correlation* between its *own* final position (once the swinging is over with) and the final position of the *pointer.* The molecule is situated in such a way as to make it possible to *infer* the final position of the pointer *from* the final position of the molecule.

(unlike the state in (5.2), which is how things stand at this same point in the story, on my friend's theory, in the *absence* of the air molecule) has *no* definite value of zip − color.[15]

And the way that things stand on my *own* theory, at this point in the story, in the presence of an air molecule, is

$$
\begin{array}{lll}
.5) & |\text{hard}\rangle_e|\text{``hard''}\rangle_m|\text{center}\rangle_a & \text{(with probability } \tfrac{1}{2}\text{)} \\[1em]
 & |\text{soft}\rangle_e|\text{``soft''}\rangle_m|\text{right}\rangle_a & \text{(with probability } \tfrac{1}{2}\text{)}
\end{array}
$$

Of course, neither of the states in (5.5) has any definite value of zip − color either. And so (5.4) and (5.5) (unlike (5.2) and (5.1)) cannot be distinguished from one another by means of measurements of zip − color.

And so, in the presence of an air molecule (unlike in the absence of one), my theory of the collapse and my friend's theory of the collapse cannot be distinguished from one another by means of measurements of zip − color.

What's going on here is that the state in (5.4) is nonseparable between the electron and the measuring instrument *and* the air molecule. Any measurement by means of which (5.4) can possibly be distinguished from (5.5)—any measurement, that is, by means of which my friend's theory can possibly be distinguished from my theory, in the presence of an air molecule (and of course such measurements will still, in principle, *exist*)—must necessarily be a measurement of an observable of the composite system consisting of all *three* of those objects. And the measurement of *that* observable (whatever, precisely, it turns out to be) will patently be a more difficult matter than the measurement of zip − color is.

[15] The way to confirm that, of course, is to write (5.4) out in terms of eigenstates of zip − color.

What's going on here, by the way, can be looked at as another special case of the very general phenomenon described in note 4 of Chapter 4. The air molecule acts here as a device for measuring the position of the pointer; and zip, and zip − color, are both incompatible with that position, and so the dynamical equations of motion themselves will entail that interaction of the air molecule with the pointer will *disrupt* the values (that is: it will *randomize* the values) of zip and zip − color.

And this is also patently not the end of the story. There will be other air molecules around too, in general, and there will be tiny specks of dust, and imperceptible rays of light, and an unmanageable multitude of other sorts of microscopic systems as well, and all of them together (in virtue of their sensitivities to the position of the pointer; in virtue, that is, of the fact that each of them can act more or less as a *measuring instrument* for the position of that pointer) will *fantastically* increase the complexity of the observables which need to be measured (which is to say: it will fantastically increase the difficulty of measuring the observables which need to be measured) in order to distinguish between my theory and my friend's theory.

And of course the business of avoiding these sorts of complications by means of perfectly isolating the pointer from any interactions whatsoever with (say) molecules of air, and rays of light will be fantastically complicated too.

The upshot of all this (that is: what these arguments *establish*) is that different conjectures about precisely where and precisely when collapses occur are the sorts of conjectures which (for all practical purposes; or, rather, for all *presently* practical purposes) cannot be empirically distinguished from one another.[16] And so apparently the

[16] What these arguments establish, by the way, has been rather widely misunderstood in the physical literature. The main confusion dates back (I think) to Daneri, Loinger, and Prosperi, 1962, and has since been carried on by Gottfried, 1966 (and there are related confusions in the work of Peres, 1980, and Gell-Mann and Hartle, 1990), and it goes like this: what these arguments establish is that different conjectures about *whether or not* collapses *ever* occur are the sorts of conjectures which (for all practical purposes; or, rather, for all presently practical purposes) cannot be empirically distinguished from one another.

Let's rehearse (again) why that can't possibly be true. The point is just that 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 (which is what all those misunderstanders always suppose), then (as we saw in Chapter 4) what proves that there *are* such things in the world as collapses is just the fact that *measurements have outcomes!* And the fact that measurements *do* have outcomes (as opposed to facts about precisely *when* they have outcomes) isn't the sort of fact that we learn by means of the difficult sorts of experiments we've been talking best we can do at present is to try to think of precisely where and precisely when collapses might *possibly* occur (precisely where and precisely when they might occur, that is, without contradicting what we *do* know to be true, as of now, by experiment).

And it turns out to be hard (as we're about to see) to do even that.

## Trying to Cook Up a Theory

Let's start over.

Let's see if we can think of precisely what it is that we *want* from a theory of the collapse.

Here's a first try at that:

i. We want it to guarantee that *measurements* (whatever, precisely, that term turns out to mean) *always have outcomes;* we want it to guarantee (that is) that there can never be any such thing in the world as a superposition of "measuring that $A$ is true" and "measuring that $B$ is true."

about *here*. The fact that measurements have outcomes is the kind of thing that we learn by means of direct introspection, by means of merely knowing that there are matters of fact about what our beliefs are!*

Of course (and this is going to be important), if it should turn out that the standard way of thinking about superpositions *isn't* the right way of thinking about them, then all bets are going to be off. But of that more later.

*It isn't easy to imagine precisely how all of those (extremely distinguished) misunderstanders can possibly have gotten themselves so confused. The argument we've been talking about, after all, is an argument about the probabilities of certain measurements having certain outcomes; any argument about the probabilities of certain measurements having certain outcomes will (needless to say) *critically depend* on the assumption that that there *are* (sooner or later) such things as the outcomes of those experiments; and so any argument about the probabilities of certain experiments having certain outcomes (supposing that the standard way of thinking about superpositions is the right way of thinking about them; which is, as I mentioned above, what all those guys *did* suppose) will critically depend on the assumption that there *are* (sooner or later) such things as *collapses;* and so (supposing what those guys supposed) any argument (like the one we've been talking about) about the probabilities of certain experiments having certain outcomes can't *possibly* be an argument to the effect that there might *not* (for all we know) be any such things as collapses; and yet that's *just* the sort of an argument that all those guys (go figure!) *took* it to be.

ii. We want it to preserve the familiar statistical connections between the outcomes of those measurements and the wave functions of the measured systems just before those measurements. That is: we want it to entail, or we want it at least to be consistent with, principle D of Chapter 2.

iii. We want it to be consistent with everything which is experimentally known to be true of the dynamics of physical systems. We want it, for example, to be consistent with the fact that isolated microscopic physical systems have never yet been observed *not* to behave in accordance with the linear dynamical equations of motion, the fact that such systems, in other words, have never yet been observed to undergo collapses.

The difficulty is in being absolutely explicit about (i).

Let's see if we can figure out how to do that. Let's try to be very practical. Let's focus somewhat more carefully on the physical mechanisms whereby the outcomes of measurements are ultimately recorded in measuring instruments.

What jumps out at you right away is that those recordings are typically stored in the *positions* of things: the positions of the tips of pointers, say, or the positions of drops of ink on scrolls of paper, or of bits of pencil lead in experimental notebooks.

Let's try to run with that. Perhaps it will suffice for what we want (which is to guarantee that measurements always have outcomes) to cook up a theory (if we can manage to) which somehow entails that every macroscopic object (that is, say, every object big enough to see, or every object even remotely big enough to see, or something like that) always (or maybe only almost always) has some definite particular position.

Let's see if we can make up a theory that does that.

Suppose we were to conjecture that every elementary particle in the world[17] occasionally (that is: once in some very great while), as

[17] That is: the indivisible elementary pointlike constituents (electrons, say, and quarks, and so on) out of which every material object in the world is presumed to be made.

The idea of such elementary particles, by the way (that is: the idea of them which we shall want to make use of here), is an explicitly *nonrelativistic* idea; and so the a matter of physical law, ceases (for an instant) to evolve in accordance with the dynamical equations of motion and undergoes a collapse which leaves it in an eigenstate of position. The times at which these collapses occur are stipulated to be absolutely random: there is merely a fixed, small, lawlike *probability,* per unit time, for each particle, that that particle, in that time interval, will undergo one of these collapses; and the point in space onto which the wave functions of these particles collapse (when they *do* collapse) will be determined probabilistically by the conventional quantum-mechanical probability formula (the one in principle D of Chapter 2); and all this is stipulated to hold for each particle *separately* (the times at which different particles undergo collapses, for example, will in general be unrelated to one another).

Let's spell this out somewhat more carefully. Consider a particle whose state at a certain instant is

6) $|A\rangle = a_1|x_1\rangle + a_2|x_2\rangle + a_3|x_3\rangle + \ldots$

There is (according to this conjecture) a certain fixed, extremely small probability that this particle will undergo a collapse within the next, say, millisecond; and if it *does* happen to undergo one of those collapses, just at the moment when its state happens to be the one in (5.6), then the *probability* that that collapse will leave it in the state $|x_1\rangle$ will be $|a_1|^2$, and the probability that that collapse will leave it in the state $|x_2\rangle$ will be $|a_2|^2$, and so on.

Let's put it slightly differently. What happens in a state like (5.6), if a collapse happens to take place, is that one of the terms in (5.6) gets multiplied by a finite number, and all the rest get multiplied by 0; and then, until the next collapse takes place (which will likely be a very long time), the state of the particle evolves again in accordance with the equations of motion. The probability, if a collapse *does* occur, that the *i*th term in (5.6) is the one that gets multiplied by a finite number is $|a_i|^2$; and the *number* it gets multi-

theory of collapses which follows is going to be an explicitly nonrelativistic *theory.* The *relativization* of this theory is something we can begin to think about, if we're still in the mood, once the *non*relativistic theory is in place.

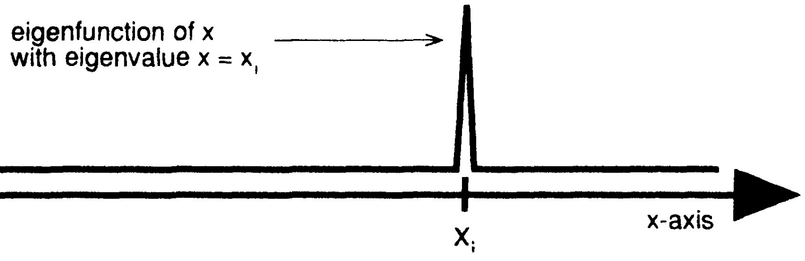

*Figure 5.2*

plied *by* (*if* it turns out to be the one that gets multiplied by a finite number) is $1/a_i$.

Let's put it one more way (this one will be particularly useful in what follows). What happens when a particle undergoes a collapse is that the *wave function* of the particle gets multiplied by an eigenfunction of the *position operator,* like the one depicted in figure 5.2; and the *probability* that the position *eigenvalue* of that position eigenfunction is $x_i$ is stipulated to be equal to $|\langle x_i|w\rangle|^2$ (where $|w\rangle$ is the state of the particle at the moment just before the collapse occurs); and note that the outcome of this multiplication (that is: the product of these two wave functions), whatever $|w\rangle$ happens to be, is invariably *also* an eigenfunction of the position operator with eigenvalue $x_i$.[18]

Suppose that the probability per minute, for each particle, of undergoing a collapse is stipulated to be very, very, *very* small (one in trillions, say). Then the probability of our ever experimentally *observing* a collapse in an isolated microscopic system, a system (that is) which consists of small numbers of particles, will be very small too (even though such collapses will necessarily sometimes occur).

But (here's the punch line) consider what happens in *macroscopic* systems; consider what happens, for example, in *measuring instru-*

[18] Note, by the way, that what I mean by *multiplying* two wave functions by one another, or taking the *product* of two wave functions, is something entirely different from taking the product of their two *state vectors.* The product of two state vectors is invariably a number. The product of two *wave functions,* on the other hand (as I'm using the term here), is defined to be that *wave function* whose value at each particular point in space is the product of the values of the two *multiplied* wave functions at that particular point in space.

*ments*. Think about the pointer in the hardness measuring device. It consists of trillions of particles (let's call them "1" and "2" and so on). The state in (5.2), for example, written out in terms of the states of those constituent particles, will look something like this:

.7) $$1/\sqrt{2}(|x_1\rangle_1|x_1\rangle_2|x_1\rangle_3 \ldots)|\text{hard}\rangle_e + 1/\sqrt{2}(|x_2\rangle_1|x_2\rangle_2|x_2\rangle_3 \ldots)|\text{soft}\rangle_e$$

where $x_1$ is the position of the pointer when it's pointing to "hard" and $x_2$ is the position of the pointer when it's pointing to "soft."[19]

Suppose that one of the particles in the pointer, any one of them, when (5.7) obtains, were suddenly to undergo a collapse (and note that the probability of *that* occurring, within even a very small fraction of a second, will be very *high*; since the pointer consists of such a gigantic number of individual particles). Consider what that will do to the state in (5.7). Suppose that it happens to be the *i*th particle in the pointer that undergoes the collapse, and suppose (for example) that that particle happens to get collapsed onto the state $|x_1\rangle_i$. What that means is that the $|x_2\rangle_i$ vector in (5.7) gets multiplied by 0; and what *that* means (since $|x_2\rangle_i$ multiplies everything *else* in the second term in (5.7)) is that when that particle undergoes that collapse, if the collapse happens to be onto $|x_1\rangle_i$ (and the probability of that is of course precisely ½), then the entire second term in (5.7) (in (5.2), that is) will vanish, and the state will turn into the first of the states in (5.1).

And so the simple and beautiful conjecture which we are now entertaining (which is originally due to Ghirardi, Rimini, and Weber, 1986, and which was later put in a particularly nice form by Bell, 1987a) entails that collapses almost never happen to iso-

[19] The fact that pointers on measuring devices typically consist of trillions of particles, by the way, is yet another of the reasons for the extraordinary difficulty of distinguishing, by means of experiments, between various different conjectures about precisely when collapses occur. The state in (5.7) (that is: the state in (5.2)) is non separable between the electron and *every one* of the constituents of the pointer; and so observables like zip – color, which distinguish between (5.2) and (5.1), must necessarily (even if the device and the electron were to be perfectly isolated from all of the rest of the world) be fantastically complicated ones; they must be observables which make reference to every individual one of all of those trillions of constituents.

lated microscopic systems. It *also* entails that states like the one in (5.2) will, almost certainly and almost *immediately*, collapse, with the standard quantum-mechanical probabilities, onto one or the other of the two states in (5.1). And all of that is of course precisely what we *want* from a theory of the collapse of the wave function; and all of it follows from a theory which can be formulated with perfect scientific explicitness, with no talk whatever (at a fundamental level) about "measurements" or "amplifications" or "recordings" or "observers" or "minds." And so now we really seem to be getting somewhere.

One technical difficulty needs to be attended to: The collapses described above leave the particles which undergo them in perfect eigenstates of the position operator, and of course that entails that the *momenta* and the *energies* of those particles (whatever their values may have been just *prior* to those collapses) will be completely uncertain just following those collapses, and *that* will give rise to a host of problems: The momenta which electrons in atoms might sometimes acquire in the course of such collapses, for example, would be enough to knock them right out of their orbits; and the energies which certain of the molecules of a gas might sometimes acquire in the course of such collapses would be enough to spontaneously *heat those gasses up*, and those sorts of things are experimentally known *not to occur.*[20]

The way to deal with that (and this is precisely what Ghirardi, Rimini, and Weber did) is to change the prescription slightly: Stipulate that when a particle undergoes a collapse, what its wave function gets multiplied by *isn't* an eigenstate of the position operator but is rather a *bell-shaped* function like the one in figure 5.3. Also stipulate that the *probability* of that bell curve's being centered at the point $x_i$ (if such a collapse happens to occur) is proportional to $|\langle B_i|w\rangle|^2$ (where $|B_i\rangle$ is the bell-curve state centered at $x_i$ and $|w\rangle$ is the state of the particle just *prior* to the collapse).

And note that in typical cases (when the wave function of the

[20] That is: they're known not to occur as *often* as would be required, statistically, by this prescription.

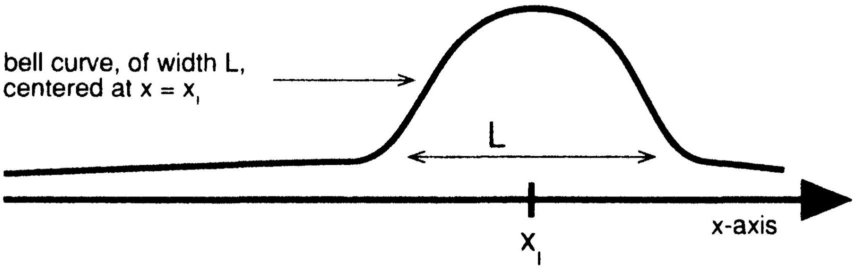

*Figure 5.3*

particle just prior to the collapse is a good deal more spread out than the bell curve, and a good deal more smoothly varying than the bell curve) the multiplication of the initial wave function by a bell curve like the one in figure 5.3 will produce a *new* wave function which is *itself* much like the bell curve in figure 5.3 and which is centered at precisely the same point.

Let's see why that helps. Remember what it is that we want from collapses. What we want from them (what will suffice, at any rate) is to make the macroscopic world look as it looks to *us:* what we want from them is to insure that macroscopic objects invariably have determinate locations, or that they almost invariably have determinate locations, or at least that they almost invariably have *almost* determinate locations. And it turns out that this revised prescription can deliver that; it turns out that the bell curves can be made narrow enough so that whatever uncertainties there are in the positions of macroscopic things are almost invariably *microscopic* ones. And it turns out (and this is the punch line) that these curves can nonetheless be made *wide* enough (at the same time) so that the violations of the conservation of energy and of momentum which the multiplications by these curves will produce will be *too small to be observed.*[21]

But as a matter of fact this revised prescription gives rise to another sort of difficulty, a somewhat more subtle one.

[21] Let's step back here a minute and look at precisely what sort of a theory we have.

This theory introduces two new fundamental constants of nature into our picture of the world: the probability, per unit time, per particle, of a collapse, and the

Consider again precisely what we want from a collapse; consider in precisely what *sense* we want it to guarantee that macroscopic objects almost always have "almost determinate locations." What we've found out is that the so-called widths of the bell curves of this prescription (that is: the length $L$ in figure 5.3) can be made microscopic; but consider whether that's what *counts,* consider whether that will completely *suffice.*

The trouble is that the values of the bell-curve functions don't ever reach precisely 0, no matter how far away from their centers you get; and so the states that particles get left in, if they undergo collapses, on the revised prescription, are still (strictly speaking) *superpositions* of being *all over the place;* and so as a matter of fact, these collapses do *not* have the effect, on the standard way of thinking, of putting *anything* in an even approximately determinate position; and so this revised prescription (on the standard way of thinking) apparently *cannot* insure (for example) that experiments with measuring devices with pointers on them ever have outcomes after all.

The effect that these collapses *do* have, of course, is to put the state vectors of pointers *close* to *other* state vectors of those pointers in which those pointers *do* have (approximately) determinate positions. What needs to be made clear (if we want to stick with this theory) is how *that* does us any good. And it isn't easy to see (so far as I know) precisely how to do that.[22]

But let's suppose that there is a way to do that, and try to press on.

*widths* (the $L$'s of figure 5.3) of the multiplying bell curves. The values of those constants, as well as the fact that such collapses occur at all, aren't things that Ghirardi, Rimini, and Weber ever make any attempt to explain: all of that is simply taken to be a part of the basic laws of nature.

This may strike some readers as unpleasantly ad hoc.

Those readers certainly have their nerve. What were they *expecting,* precisely? Let them go and reflect on what came *before;* let them go and reflect how much has been accomplished here.

[22] Doing that (if there *is* a way of doing that) will apparently require some sort of a modification of the standard way of thinking.

## Experiments with Television Screens

What we've been supposing so far is that every measuring instrument must necessarily include some sort of *pointer,* which indicates the outcome of the measurement. And we've been supposing that that pointer (if this instrument really deserves to be called a measuring instrument) must necessarily be a macroscopic physical object. And we've been supposing that that pointer must necessarily assume macroscopically different *spatial positions* in order to indicate different such outcomes. And it turns out that if all that is true, then the GRW theory (which is what we'll call it from now on) can do (i) and (ii) and probably (iii) as well.

The question, of course, is going to be whether all measuring instruments (or, rather, whether all reasonably *imaginable* measuring instruments) really do work like that.

Here's a standard sort of arrangement for measuring the hardness a particle: The particle gets fed into a hardness box like the one in figure 5.4, which separates hard particles from soft ones by means of magnetic fields. The incoming hard particles get directed (by the magnetic fields) toward point *A* on a TV screen, and the incoming soft particles get directed toward point *B*. The TV screen works like this: A particle striking the screen at (say) point *B* knocks

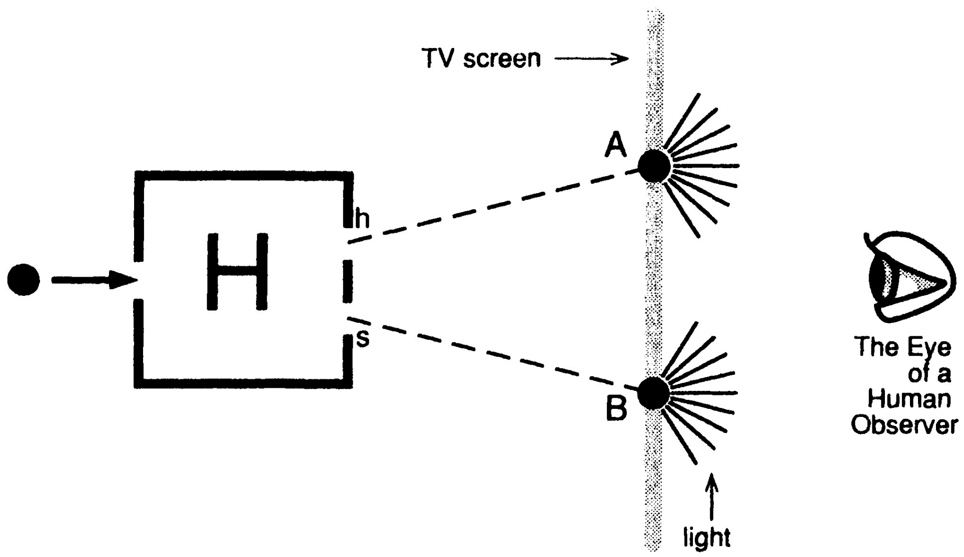

*Figure 5.4*

certain electrons in certain of the fluorescent atoms in the screen in the vicinity of point *B* into their so-called excited orbits (that is: it knocks those electrons into orbits around their atomic nuclei which are a good deal more *energetic* and a good deal more *unstable* than the orbits they're normally in); and soon thereafter those electrons spontaneously de-excite back to their original (stable, low-energy) orbits; and in the process of de-exciting (that is: in the process of losing their excess energy) they emit photons; and thus the vicinity of *B* becomes a luminous dot, which can be observed directly (that is: without any further artificial apparatus) by a human experimenter.

We want to inquire whether or not the GRW theory entails that a hardness measurement like that, on a particle which is initially (say) black, has an outcome. That will depend on whether or not there ever necessarily comes a time, in the course of measurement, when the position of a macroscopic object, or the positions of some gigantic collection of microscopic objects, is *correlated* with the hardness of that particle (let's call it *P*). With all that in mind, let's rehearse the stages of the measuring process (what follows here gets spelled out in a bit more detail in Albert and Vaidman, 1988) again.

First, the wave function of *P* gets magnetically separated (by the hardness box) into hard and soft components. No outcome of the hardness measurement (no collapse, that is) will be precipitated by that separation, since, as yet, nothing in the world save the position of *P* (nothing, that is, save a single, microscopic degree of freedom) is correlated to the hardness. Let's keep looking.

Next, *P* hits the screen, and at that stage the electrons in the fluorescent atoms get involved. Consider, however, whether those electrons get involved in such a way as to precipitate (via GRW) an outcome of the hardness measurement. Here's the crucial point: the GRW "collapses" are invariably collapses onto (nearly) eigenstates of position, but it's the *energies* of the fluorescent electrons, and *not* their positions, that get correlated, here, to the hardness of *P*! The GRW collapses aren't the right *sorts* of collapses to precipitate an outcome of the hardness measurement here.

Let's make this point somewhat more precise. If the initial state of *P* is $|\text{black}\rangle_P$, then, just after the impact of *P* on the TV screen, the (nonseparable) state of $P$ and of the various fluorescent electrons in the vicinities of $A$ and $B$ will look (approximately; ideally) like this:

$$\begin{aligned} &1/\sqrt{2}\ |\text{hard}, X = A\rangle_P |\text{ex}\rangle_{e1} \ldots |\text{ex}\rangle_{eN} |\text{unex}\rangle_{eN+1} \ldots |\text{unex}\rangle_{e2N} \\ &\quad + 1/\sqrt{2}\ |\text{soft}, X = B\rangle_P |\text{unex}\rangle_{e1} \ldots |\text{unex}\rangle_{eN} |\text{ex}\rangle_{eN+1} \\ &\quad \ldots |\text{ex}\rangle_{e2N} \end{aligned} \tag{.8}$$

where $e1 \ldots eN$ are fluorescent electrons in the vicinity of $A$, $eN + 1 \ldots e2N$ are fluorescent electrons in the vicinity of $B$, $|\text{ex}\rangle$ represents a state of being in an "excited" orbit, and $|\text{unex}\rangle$ represents a state of being in an "unexcited" orbit. Suppose, now, that a GRW "collapse" (that is, a multiplication of the wave function of one of the fluorescent electrons by a bell curve like the one depicted in figure 5.3) occurs. Consider whether this sort of a collapse will make one of the terms in (5.8) go away, allowing only the other to propagate. The problem, once again, is that these aren't the right sorts of collapses for that job, because $|\text{ex}\rangle$ can't be distinguished from $|\text{unex}\rangle$ in terms of the *position* of anything. Indeed, a GRW collapse will leave (5.8) almost entirely unchanged (except, perhaps, in the wave function of some single one of the many, many fluorescent electrons). And so no outcome of this measurement is going to emerge at this stage of things, either.[23]

We shall have to look still elsewhere. The next stage of the measuring process involves the decay of the excited electronic orbits and (in that process) the emission of photons. If the first term of (5.8) obtains, the photons would be emitted at $A$; if the second term obtains, the photons would be emitted at $B$. Those two states, then, *can* be distinguished, at least at the moment of emission, in terms of the *positions* of the *photons*. Now, so far, the GRW theory has been explicitly written down in a form which applies only to

[23] I've left aside the whole question of the *probability* of a GRW collapse taking place at this stage of things (since, as we've just seen, a collapse like that won't do any good here anyway), but it ought to be noted in passing that that probability might well turn out to be extremely low. It's well known, after all, that the unaided human eye is capable of detecting very small numbers of photons; so perhaps only very small numbers of fluorescent electrons need, in principle, be involved here!

*non*relativistic quantum mechanics. The behaviors of *photons,* on the other hand (and particularly the *creation* of photons, by devices like TV screens), can be accounted for only in the context of a *relativistic quantum field theory.* It isn't completely clear as yet (for a number of rather technical reasons) how a GRW-type theory might treat *them.* If it turns out that photons can't experience GRW collapses, then of course no outcome of the hardness measurement can possibly emerge at this stage. But let's give the theory the benefit of the doubt: let's suppose that photons *can* experience GRW collapses. The problem at this stage of the measurement is going to be that that distinguishability is going to be extremely short-lived. It turns out that in almost no time, in far too little a time for a GRW collapse to be likely to occur (supposing that *A* and *B* are, say, a few centimeters apart, on a flat screen), the two photon wave functions described above will spread out (as shown in figure 5.5) so as to overlap almost entirely in position space, and the distinguishability in terms of positions will go away, and we shall be in just such a predicament as we found ourselves in at the previous stage of the measurement. No outcome, it seems, will emerge here, either.

But now we're running out of stages. The measurement (according to all the conventional wisdom about measurements) is already over! By now, after all, we have a recording; by now genuinely macroscopic changes (that is: changes which are thermodynamically irreversible, changes which are directly visible to the unaided human eye) have already taken place in the measuring apparatus.

And so it turns out that genuine recordings need not entail macroscopic changes in the position of anything; changes in (say) the energies of large numbers of atomic electrons (as in the above example) can be recordings too. It turns out (to put it slightly differently) that there can be genuinely macroscopic measuring instruments that (nonetheless) have absolutely no macroscopic *moving parts.*

That's what's been overlooked in the GRW proposal. What the GRW theory requires in order to produce an outcome isn't merely that the recording in the measuring apparatus be macroscopic (in any or all of the senses just described), but rather that the recording

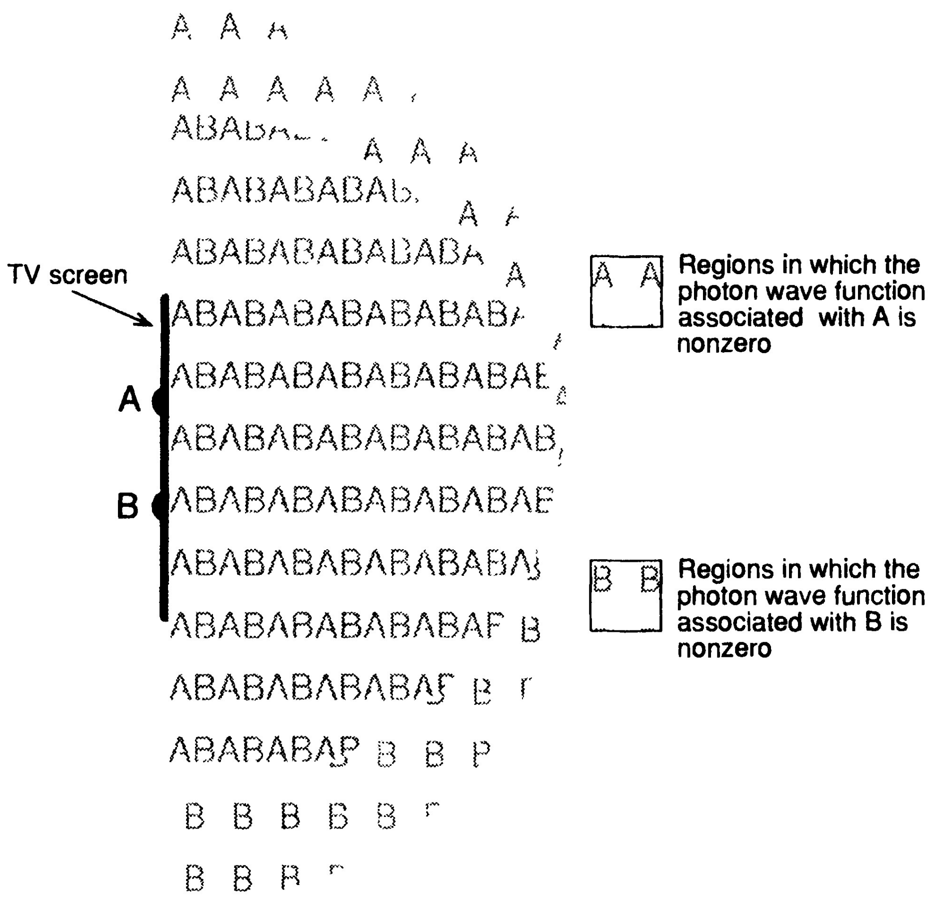
TV screen
A
B
Regions in which the photon wave function associated with A is nonzero
Regions in which the photon wave function associated with B is nonzero

*Figure 5.5*

process involve macroscopic changes in the *position* of something. And the trouble is that no changes of that latter sort are involved in the kinds of measurements we've just been talking about.[24]

## Inside the Observer's Head

Suppose, after all this, that we wanted to stick with the GRW theory anyway. What would that entail?

Well, we would have to deny that the measurement described above is over even once a recording exists. We would have to insist

[24] Needless to say, this is not a point merely about measuring devices. The situation is as follows: if the GRW theory is the true physical theory of the world, then (astonishingly) things can in principle be cooked up in such a way as to guarantee that there literally fails to be any matter of fact about whether (say) Ralph Kramden's face or Ed Norton's face is the face that appears on some particular TV screen, at some particular moment.

(and certainly this is an ineluctable fact, when you come right down to it) that no measurement is absolutely over, no measurement absolutely requires an outcome, until there is a *sentient observer* who is actually *aware* of that outcome.

So, if we wanted to try to stick with this theory in spite of everything, the thing to do would be to insist that as a matter of fact we haven't run completely out of stages yet, and to go on looking, in those latter stages, for an outcome of this experiment (even though we've already looked right up to the retina of the observer and not found one); and of course the only place *left* to look at this point is going to be on the inside of the *nervous system* of the observer.

This is going to be an uncomfortable position to be in; what the possibility of entertaining some particular fundamental theory of the whole physical world is going to hinge on (if we go down this road) are the answers to certain detailed questions about the physiology of human beings; but let's try to press on and see how it might work. Here's one idea: Consider the two different physical states of the observer's retina (call them *RA* and *RB*) which arise in consequence of its being struck by light from one or the other of the two luminous spots on the fluorescent screen. Perhaps the retinal states *RA* and *RB* macroscopically differ from one another in (among other things) the *positions* of some gigantic collection of microscopic physical objects (the positions of some collection of ions, say). Whether or not that turns out to be so is of course a question for neurophysiology (and I presume that it's a question for the future of that subject), but the idea would be that if it *does* turn out like that, then the *observer's retina itself* can play the role of GRW's macroscopic pointer in this experiment: *it* (the retina) can bring about the collapse, *it* can suffice to finally precipitate an outcome.

Now of course it might turn out that *RA* and *RB* do *not* differ from one another in the positions of any sufficiently gigantic collection of physical objects. It might turn out, say, that the retina works more or less like a fluorescent screen and (consequently) that no outcome of this experiment can emerge at the stage of the retina, either. In that case we would presumably turn our attention next to the optic nerve, and *then*, finally (if things go badly with the optic nerve too), to the observer's brain itself.

If things were ever to get to that point, then the possibility of continuing to entertain the GRW theory would hinge directly on whether or not the brain state associated with seeing a luminous dot on the fluorescent screen at point *A* (call that *BA*) differs from the brain state associated with seeing a luminous dot on the fluorescent screen at point *B* (call that *BB*) in terms of the positions of any gigantic collection of physical components of the observer's visual cortex. If those two brain states *do* differ in that way (and that, once again, will be a question for neurophysiology), then the GRW theory *could* continue to be entertained.

But of course if it were to emerge, after all this, that *BA* and *BB* do not differ in precisely that way, if it were to emerge that they differ only in ways other than that, then the game would be over; then (that is) we would be absolutely out of stages.

Suppose that the human neurophysiology works out O.K.; suppose (that is) that it turns out that the human brain states *BA* and *BB* which I just described (the states which correspond to seeing a luminous dot at point *A* and seeing a luminous dot at point *B*) differ by, among other things, the *positions* of some gigantic number of ions.[25] How would things stand then?

Well, that would be a relief. That would mean that the GRW theory entails that at the point when the visual cortex of the experimenter gets into the game, then (and *only* then) an outcome of this experiment does finally emerge. Of course this outcome comes astonishingly late (later than we can possibly be comfortable with, later than anyone could ever have imagined). But suppose we're willing to swallow all that; suppose (that is) that we're willing to entertain the possibility that our intuitions can be radically false even about the absolutely familiar macroscopic external world. Then it would have to be admitted that (if you go strictly by the rule book) this outcome comes on time.

[25] As a matter of fact, it has recently been argued (Aicardi, Borsellino, Ghirardi, and Grassi, 1991) that the brain states produced by different visual stimuli *are* likely to differ in that way. It isn't clear yet, however, what the story is with other sorts of sensory stimuli.

But what if at that point we were to begin to worry about the possibility of there being *other* sorts of sentient observers in the world (dolphins, say, or Martians or androids or whatever) for whom *BA* and *BB* *don't* differ in the appropriate way?

Or what if we were to begin to worry about the possibility of the development of (say) surgical techniques whereby ordinary human beings could be *transformed* into beings for whom *BA* and *BB* don't differ in the appropriate way?

Let's work through a science-fiction story (a story which, however, is everywhere scrupulously consistent with the hypothesis that the GRW theory is the true and complete theory of the world) about that last possibility.

First we'll need to set things up. The story involves a device (figure 5.6) for producing a correlation between the position of a certain microscopic particle *P* and the hardness of an electron; a sort of measuring device for hardness, in which the "pointer" is this particle called *P*. Here's how the device works: If *P* starts out in its middle position, and if a hard electron is fed through the device, then the hardness of that electron is unaffected by its passage through the device, and the device is unaffected too, except that *P* ends up, once the electron has passed all the way through, in its upper position; and if *P* starts in its middle position, and if a *soft* electron is fed through the device, then the hardness of that electron is unaffected by its passage through the device, and the device is unaffected too, except that *P* ends up in its *lower* position.

Suppose now that sometime in the distant future somebody named John undergoes a technically astonishing neurosurgical pro-

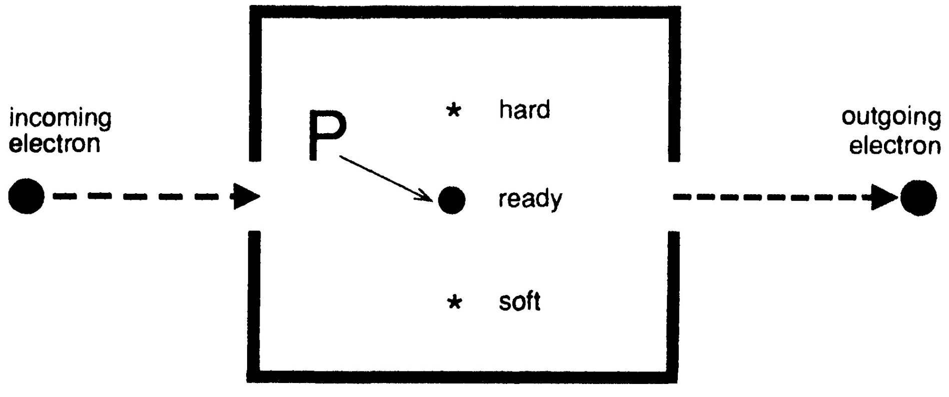

*Figure 5.6*

cedure which leaves him looking like he does in figure 5.7: John has a little door on either side of his head, and a device like the one I just described is now sitting in the middle of his brain, and (here comes the crucial point) the particular way in which that device is now hooked up to the rest of John's nervous system makes John behave as if his *occurrent beliefs* about the hardnesses of electrons which happen to pass through that device are determined *directly* by the position of *P*.

Here's what I mean. Suppose that John is presented with a hard electron and is requested to ascertain what the value of the hardness of that electron *is*. What he does is to take the electron into his head through his right door, and pass it through his surgically implanted device (with *P* initially in its middle position) and then expel it from his head through his left door. And when that's all done (when *P* is in its upper position but when, as yet, the value of

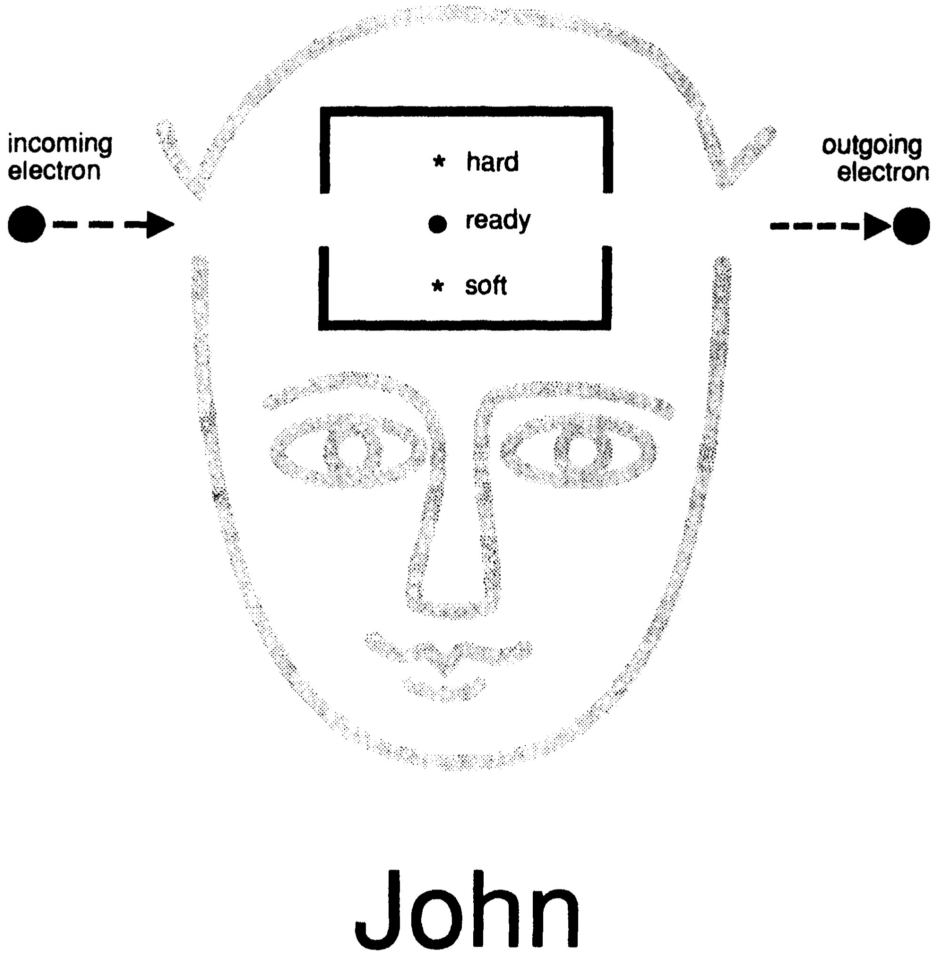

*Figure 5.7*

the hardness of the electron isn't recorded anywhere in John's brain *other* than in that position of *P*), John announces that he is, at present, consciously aware (as vividly and as completely as he is now aware of anything, or has ever been aware of anything in his life, he swears) of what the value of the hardness of the electron is, and that he would be delighted to tell *us* what that value is, if we would like to know.[26]

Let's think about what John says. There's going to be a temptation to suppose that John must somehow be mistaken, to say: "Look, how can it *be* that John is now consciously aware of what the value of the hardness of the electron is? Nothing in John's brain, other than the position of *P*, is *correlated* to the hardness of the electron now; and *P* isn't a natural part of John's brain *at all; P* is just a *surgical implant, P* is (after all) just a *particle!*" And imagine that John will say: "Look; whether or not *P* happens to be one of the components of my brain that I was born with is completely beside the point; I tell you now, from my own introspective experience, that if (as you tell me) nothing is now correlated to the hardness of that electron other than the position of *P*, then things must now be wired up in such a way that the position of *P* itself directly determines my present occurrent conscious beliefs about the value of the hardness of that electron!"

And John is now indeed in a position to announce, correctly, what the value of the hardness of that electron is, if he's asked to; and it's pretty clear that John can in principle be wired up so as to reproduce, in his present state, *any* of the behaviors of a genuine "knower" of that hardness *whatsoever;*[27] and John will claim (and who will we be to argue with him?) that, after all, no one is better

[26] Of course, in the event that John actually *does* tell us that value, then (in the course of the physical process of telling us that) certain physical observables of John's brain other than the position of *P* would, no doubt, become correlated with the hardness of the measured electron. The point is that at *this* stage of things (when it's already the case that John claims to *know* the value of that hardness) they *aren't.*

[27] And it's also pretty clear that John can in principle be wired up in such a way as to guarantee that (in his present state) he satisfies any *functional* criterion you can dream up for being a genuine "knower" of the hardness.

qualified than he himself to judge whether his own psychological experiences are really "genuine" ones or not!

Anyway, this much is certain: Either John is radically mistaken about what his own psychological experiences are, or else (if he isn't mistaken) nothing like a GRW theory can possibly insure that there are invariably determinate matters of fact about the psychological experiences of sentient observers at all. The point is that nothing macroscopic has happened in this story yet; and so if John now actually *has* an occurrent belief about the hardness of the electron, as he says he does, and as everything that can be empirically found out about him testifies he does, then John can indeed come to have such beliefs without anything macroscopic happening; and so (in the event that, say, the electron whose hardness John measures starts out in an eigenstate of *color*) nothing along the lines of a GRW theory is going to be able preclude the development of a *superposition* of states corresponding to *different* such beliefs.

And as a matter of fact, if John *can* come to have beliefs without anything macroscopic happening (which he says he can do, and which it seems he can do), then *no* theory of the collapse of the wave function *whatsoever* is going to be able to preclude the development of superpositions like that; because precluding superpositions like *that* will require that the collapse be inserted at a level (the level of *isolated microscopic systems*) at which we know, by experiment, that no collapses ever occur.[28]

Of course (on top of everything we've just been talking about) there hasn't ever been so much as a shred of what you might call *normal*

[28] Perhaps this ought to be fleshed out a bit. Suppose that we were able to cook up a theory of the collapse which is patently as good as any theory of the collapse could imaginably be. Suppose, that is, that we were able to cook up a theory which is consistent with everything we know to be true of the behaviors of isolated microscopic systems and which entails (somehow) that superpositions of states which differ from one another in terms of anything macroscopic whatever (not just in terms of the *positions* of any gigantic number of particles) don't last long. And suppose that it were clear that this theory can indeed guarantee that an experiment carried out by an ordinary human observer invariably has an outcome.

The point is that a theory like that would nonetheless obviously fail to guarantee that for *John* in exactly the same way as the GRW theory does.

*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*.