# Uncertain Decisions and the Many-Minds Interpretation of Quantum Mechanics

## *1. Introduction*

Imagine you are faced with a quantum mechanical device which will display either H or T ("heads," "tails") when it is operated ("spun"). You know that the single-case probability, or *chance*, of H is 0.8, and the chance of T is 0.2. (I could have taken a biased coin, but I wanted to make it clear that we are dealing with chances.)

You are required to bet at evens (your £1 to your opponent's £1, winner take all) but have a choice of whether to back H or T. What should you do?

Back H, of course. The expected gain of this option is £0.60 ((0.8 × £1) + (0.2 × -£1)), whereas backing T has an expected gain of -£0.60 (i.e., an expected loss).

But *why* is this the right choice? This looks like a silly question. Haven't I just shown that backing H will probably make you money, whereas backing T will probably lose you money? But we need to go slowly here. Note that any normal person wants *to gain money*, not *probably* to gain money. After all, what good is the bet that will probably make you money, in those cases where the device displays T and you lose a pound? So, once more, why should you choose the option that will *probably* make you money?

Even if this question looks silly at first, it cuts very deep. In this paper I shall proceed as follows. In Section 2, I shall show that there is no good answer to my question, and that this is puzzling. In the following section I shall show there is no good answer to a related question, and that this is even more puzzling. I shall then describe the "many-minds" interpretation of quantum mechanics, and show that my puzzles simply dis-

"Uncertain Decisions and the Many-Minds Interpretation of Quantum Mechanics" by David Papineau, *The Monist*, vol. 80, no. 1, pp. 97–117. 

appear if we accept the many-minds view. Where this leaves us I am not sure. The many-minds theory is so counter-intuitive that it is difficult to take seriously. But it does have independent motivation in the philosophy of physics. Given this, it is at least noteworthy that it also dissolves one of the most deep-seated and difficult conundrums in the philosophy of probability and action.

*2. Probabilities and Pay-Offs*

My initial question was: why we should do what will probably get us what we want? To help clarify this request for an explanation, compare the following two standards by which we might assess which of a range of options is prudentially the best (that is, the best way for you to get what you want).

(A) The prudentially best option is the one which will in fact yield the greatest gain. (Call such options "best$_A$".)

(B) The prudentially best option is the one which maximizes your probabilistically expected gain.[1] (Call these options "best$_B$".)

Now, there seems an obvious sense in which (A) gives us the fundamental sense of prudentially "best." From the point of view of your getting what you want, the action which in fact yields the greatest gain is surely the best.

Given this, we can understand my initial question as a request for an explanation of (B) in terms of (A). On the assumption that the best$_A$ actions are primitively best, can we find some sense in which best$_B$ actions are derivatively best?

The initial difficulty facing any answer is that there is no guarantee that any particular best $_B$ option will also be best$_A$. You bet on H, as this is certainly best$_B$. But this choice will not be best$_A$ in those improbable (but all too possible) cases where the device displays T. So, once more, what is the worth of best$_B$ choices, given that they won't always be best$_A$?

Some readers might wish to object here that we often have no alternative to best$_B$ choices. In cases like our original example, we don't know what is actually going to happen. We only know the chances of the alternative outcomes. Given that practical choices in these cases *can't* be guided by

(A), isn't it obvious that we should use (B) instead? If we don't know which action will actually succeed, how can we do better than choose the action that will most probably succeed?

But this doesn't answer. It doesn't connect (B) with (A). We haven't yet been told why the action that will *probably* get us what we want is a good way of aiming at what we *actually* want. Let us grant that we need *some* strategy for choosing actions when we can't use (A). Still, what is so good about (B)? Why not instead opt for the action which *minimizes* the positive expected gain, rather than maximizes it? This action might still give you what you want. Of course, it is less *likely* to do so than the $\text{best}_B$ action. But this last thought does not explain (B), so much as simply restate it.

Can't we offer a *long-run* justification for (B) in terms of (A)? Won't repeatedly choosing according to (B) in situations of uncertainty yield greater actual gain in the long run than any other strategy? However, there are two objections to this argument.

(1) First, choosing according to (B) *won't* necessarily lead to the greatest gain in a sequence of repeated choices. Suppose you repeatedly back H in our earlier example. This won't be the most lucrative strategy if you suffer an unlucky sequence of mostly Ts. In that case repeatedly backing T would have done better. (What about the *infinite* long-run? Won't (B) be guaranteed to maximize gains in an *infinite* sequence of repeated choices? Maybe so, if you buy the frequency theory of probability, but this resort to infinity only adds weight to the second objection.)

(2) The truth of (B) seems quite independent of any long run considerations (and *a fortiori* independent of infinite long-run considerations). Let us agree that there is some sense (yet to be pinned down) in which the best option for an uncertain agent is indeed the $\text{best}_B$ option that maximizes expected gain. The advisability of this option surely doesn't hinge on what gains similar choices may or may not bring in the future. Suppose I expect the world to end tomorrow. Or, more realistically, suppose that I am a shiftless character with little thought for the future. I want some money *now*, and who cares what tomorrow will bring. Surely I have just as much reason as anybody else to bet with the odds

> rather than against. From the point of view of my desire for money now, the right thing is clearly to back the 80% H, and not the 20% T. My lack of interest in future bets surely doesn't lessen the advisability of my backing H in this case.

It seems that, whichever way we turn it, we cannot derive (B) from (A). In recognizing this, I am merely following a long line of philosophers. At first sight, it seems that we ought to be able to explain why it is good to act on probabilities. But there is no non-question-begging way of showing that the option which will most probably succeed is a good way of aiming for what you actually want. So philosophers from Peirce (1924) to Putnam (1987) have felt compelled to conclude that principle (B) is a brute fact about rational choice, which lacks further explanation. We should stop asking *why* it is right to act on probabilities. This is just what it means to choose rationally in conditions of uncertainty.

Perhaps we should learn to live with this. Our spade must turn somewhere. But I find it very odd that we should have two independent principles for assessing choices. True, this independence engenders no practical conflict. For, as I observed earlier, in situations of probabilistic uncertainty there is no possibility of using (A), and so we simply use (B) instead. Even so, (A) still *applies* to uncertain decisions, and seems to give the fundamental sense of the prudentially best option. So it seems very odd that we should actually make uncertain choices *via* some quite independent standard for assessing actions, namely, (B).

I want to explore a different path. So far we have started with (A) and tried to explain (B). What if we start with (B) instead, and take it to be fundamental that the best actions are those that maximize expected gain? Can we then explain (A) in terms of (B)?

The immediate problem facing this strategy is that the $\text{best}_B$ action sometimes isn't $\text{best}_A$, as when backing H turns out badly because of an unlucky T. But let us put this difficulty to one side for the moment, and note that there is one respect in which the (B)-first strategy works better than the earlier (A)-first strategy. If we take (B) as basic, then we don't have to invoke any other principle to explain all the prudential choices we actually make.

This is because (A) is only a practical guide to action in conditions of certainty. But in conditions of certainty (A) is a special case of (B).

When agents know which outcomes will occur, then the chances of those outcomes are already fixed as 0 or 1, and the $best_B$ choice which maximizes probabilistically expected gain is identical with the $best_A$ choice which in fact yields the greatest gain.

This doesn't work the other way round. When we took (A) to be basic, we could of course immediately account for rational choices made under conditions of certainty. But (A) didn't help at all with choices made under uncertainty. The $best_B$ choices that maximize probabilistically expected gains certainly aren't always special cases of the $best_A$ choices that actually get you what you want, nor is it clear how they relate to them. This was the basic problem facing (A)-first strategy. So, when we started with (A), we were forced to invoke (B) as another primitive principle, to account for uncertain choices.

I think this gives us some reason to suspect that (B), rather than (A), encapsulates the primitive truth about prudentially best choices. If we start with (B) we have one principle that explains all our choices. Whereas, if we start with (A), we can explain choices made under certainty, but then have to invoke another unexplained principle to account for uncertain choices.

Still, what about (A)'s assessments of decisions made under uncertainty, as in our basic betting example? As I said, (B) clearly doesn't explain these assessments, since the $best_B$ choice can diverge from the $best_A$ choice in such cases. Maybe (A) is of no practical use in such cases, which is why (B) alone is able to explain all our practical choices. But (A) still *applies* in these cases, and tells us that the $best_B$ action isn't always the $best_A$ action.

Given this, has anything really been gained by "starting" with (B)? We can say (B) gives us the "fundamental" sense of "best", if we like, but don't we still face the problem of explaining what's so good about $best_B$ actions, if they don't always give us what we want?

I agree that the underlying problem hasn't really gone away if we still accept that (A) applies to decisions made under uncertainty. However, there is a more radical option. Suppose we just ditch (A) altogether. We simply deny that it is a virtue in an action that it yields what you want. The only virtue in an action is that it will *probably* yield what you want. If we do this, then we will have no problem of squaring one standard of goodness with another, for we will only have one standard of goodness left.

At first sight this option scarcely makes sense. Surely there is *some* sense in which the action which gives you the greatest gain is "best". What could it *mean* to deny this? At this stage I can only ask readers to give me the benefit of the doubt. In due course I shall describe a metaphysical picture which will make my radical suggestion comprehensible, indeed mandatory. But for the moment I only want to make the abstract point that, *if* we could somehow ditch (A), then there would be no remaining problem of squaring (B)-assessments with (A)-assessments, and moreover (B) alone would explain all our practical choices.

### *3. Single-Case and Knowledge-Relative Probabilities*

I have just promised that in due course I will make good the counter-intuitive suggestion that the only standard of goodness for choices is that they will probably yield what you want, not that they will actually do so. In fact it's even worse than that.

So far I have been assuming that uncertain agents at least know the chances of relevant outcomes. This assumption was built into in principle (B), in that I defined "maximum expected gain" as the average of gains weighted by the chances of the relevant outcomes. However, if we focus on *chances* in this way, then there will be yet another species of prudentially optimality we can't account for, a species of optimality which often guides agents who don't even know the chances. So in this section I am going to argue that it is not even principle (B) that should provide the sole standard of optimal choice, but an even more surprising principle (C).

An example will help readers to see where I am heading. Imagine a machine that makes two kinds of quantum-mechanical devices. When operated ("spun") each device displays either H or T. The first kind of device is H-biased, giving H a 0.9 chance and T a 0.1 chance, say. The other kind is T-biased, giving H a 0.3 chance and T a 0.7 chance. In addition, let us suppose that the machine that makes the devices itself proceeds on a quantum-mechanical basis, and that on any occasion there is a 0.6 chance of an H-biased device, and a 0.4 chance of a T-biased one.

Now imagine that you are faced with a particular device; you know it is from the machine, but have no indication of which kind it is. You are offered the same bet as before: £1 at evens, with you to choose which result to back. What should you do?

Back H, obviously. There's a 0.66 probability of H ($0.6 \times 0.9 + 0.4 \times 0.3$), and so a 0.34 probability of T, which means that backing H offers

an expected gain of £0.32, while backing T offers an expected loss of £0.32.

However, note that this choice is *not* necessarily optimal according to principle (B). Principle (B) works with *chances*, yet the 0.66 and 0.34 probabilities which entered into the above expected-gain calculation are not chances. The chance of H is either 0.9 or 0.3, depending on which kind of device you have, not 0.66. Moreover, if the chance of H is 0.3 (you have a T-biased device), then backing H is not the $\text{best}_B$ option, but backing T.

What kind of probability is the 0.66 probability of H, if it is not a chance? It is what I shall call a "knowledge-relative probability". This is a probability in the sense: how often does H tend to occur in situations-like-the-one-you-now-know-yourself-to-be-in: for example, how often does H occur when you spin a device from the machine? (Note that while it is subjective *which* knowledge-relative probability bears on some particular choice, since this depends on the agent's current information, such probabilities are themselves objective: for example, it is independent of anybody's opinion that H tends to be displayed in 66% of the spins of devices from the machine.[2])

This notion of knowledge-relative probability allows us to distinguish the following two principles for identifying prudentially optimal choices:

(B) The prudentially best option is the one which maximizes the *single-case* expected gain. (Call such options "$\text{best}_B$".)

(C) The prudentially best option is the one which maximizes the *knowledge-relative* expected gain.[3] (Call these "$\text{best}_C$".)

If we compare these two principles for assessing choices, analogies of all the arguments from the last section present themselves.

In our current example you do not know the single-case probability of H, but only the knowledge-relative probability. So you have no choice but to assess your options by principle (C). But from a God's-eye point of view, so to speak, principle (B) still *applies* to your choice, and moreover it seems to deliver a more basic sense of best option. After all, what you *want* is surely the option which is $\text{best}_B$. You want to back H if you've got an H-biased device, but not if you've got the other kind. In practice, of course, you'll go for the $\text{best}_C$ option, and back H, for you don't know

which option is $best_B$. But you would clearly *prefer* to be able to decide by applying standard (B) rather than standard (C).

Given this, it seems natural to try to justify principle (C) in terms of principle (B): we feel that we ought to be able to show, on the assumption that $best_B$ actions are primitively best, that $best_C$ options are in some sense derivatively best.

However, this can't be done. The initial hurdle is that the $best_C$ option simply isn't always $best_B$. (In our example, backing H won't always maximize single-case expectation.) Nor does it help to appeal to the long run. The $best_C$ choices are only guaranteed to maximize single-case expectation in the *infinite* long run (and even then only if you assume the frequency theory). Moreover, it seems wrong to argue that the rationality of *now* choosing the $best_C$ option depends on how similar choices will fare in future.

So, if we take (B) as basic, we seem forced to postulate (C) as an additional and underived principle which informs our choices in cases where we know the knowledge-relative but not the single-case probabilities. This is cogent, but odd. Principle (B) still applies in these cases, and seems to involve a more basic sense of prudential worth. Yet we decide how to act *via* a quite unconnected standard for assessing actions, namely, (C).

What if we start with (C) instead of (B)? Obviously principle (C) can't explain *all* B-assessments, for as I have just observed, the $best_C$ option isn't always $best_B$. But it does at least account for all the B-evaluations we make in cases where we *use* (B) to guide our actions. For we only use (B) when we know the single-case probabilities. And when we know the single-case probabilities (B) is a special case of (C): for example, if you know the single-case probability is 0.8, then you know that H happens in 80% of cases-like-the-situation-you-know-yourself-to-be-in. (Note that principle (A) is therefore also a special case of (C) in all those cases where *it* informs practical choices; for we have already noted that it is a special case of (B) in those cases.)

So perhaps we ought to take (C) as basic, rather than (B) or (A). However, there remains the fact that when we *don't* know the single-case probabilities, and so must act on (C), rather than (B), the $best_C$ option isn't always the $best_B$. So isn't the underlying problem still there? What's so good about the $best_C$ option, if it is all too possible that this option doesn't maximize single-case expectation?

However, consider the radical step of ditching (B) altogether (along with (A)). Then principle (C) would be the *only* standard for assessing actions; the only sense in which an action can be good would be that it maximizes knowledge-relative expectation. This may make little sense as yet. But for the moment simply note that (C) by itself would still be able to explain all the practical choices we ever make. And there would be no further problem of squaring (C)-assessments with other standards for assessing choices, if we had jettisoned the other standards.

In the next section I shall begin to sketch a metaphysical picture which will make this radical option comprehensible. But first it will be worth noting that, even if (C) is our only principle, we can still do a kind of justice to the thought that it is good to act on knowledge of the single-case probabilities, if you can. Suppose that, as before, you are faced with a device from the machine, but don't know which kind, and that you have the choice of backing H or T at evens. But now suppose that you have a third, further option W (for "wait-and-see"): you are allowed to find out which kind of device it is, at no cost (you can turn it over and see if it is marked "h-series'" or "t-series", say), and then you can decide whether to back H or T at evens. Given this three-way choice, W is clearly the best option. Where backing H immediately offers an expected gain of £0.32, and backing T immediately offers an expected loss of £0.32, W offers an expected gain of £0.64 (To see this, note that there is a 0.6 probability that the device is "h-series", in which case you will then back H and expect to win £0.80; and a 0.4 probability it is "t-series", in which case you will back T and expect to win £0.40; the weighted average over the two possibilities shows that W offers an expected gain of £0.64.)

This result is quite general. An argument due to I. J. Good (following F. P. Ramsey) shows that the option of gettng-more-free-information-and-then-betting-on-your-new-knowledge-relative-probabilities always offers at least as great an expectation as any immediate bet. The general result applies to the discovery of probabilities relative to *any* extra information, and not just to the discovery of single-case probabilities. But since finding out single-case probabilities is a special case of finding out new knowledge-relative probabilities, the result applies in these special cases too (as in the above example).

This Ramsey-Good result thus shows that it is always better to find out the single-case probabilities and act on them, if you can (and the extra information is free), rather than to act on merely knowledge-relative prob-

abilities. You might be tempted at this point to read this result as an argument for reinstating (B). Doesn't the Ramsey-Good result show that the best choice is the one informed by the single-case probabilities? But I think we should resist this reading. (After all, we don't *want* to reinstate (B) alongside (C), since this would also reinstate the awkward divergence between many (C)-assessments and (B)-assessments.)

If we consider the matter carefully, we can see that the Ramsey-Good result is entirely consistent with the complete rejection of (B) in favour of (C). To see this, note that the Ramsey-Good result uses (C), not (B), to show that W is better than any immediate bet: for it shows that W is better *on average* across the 0.6 probable "h-series" and 0.4 probable "t-series", and, since your device has already been made, these numbers are knowledge-relative probabilities, not chances. True, the proof also assumes that after you get the extra information you will bet in line with the single-case probabilities; but this too need only assume (C), since, once you know the single-case probabilities, then acting on them is a special case of conformity to (C).

I know it is hard to get one's head around the idea that the single-case probabilities are a better guide to action than knowledge-relative probabilities because allowing yourself so to be guided will maximize your *knowledge-relative* expected gain. Once more, I can only ask readers to bear with me. All will become clear in due course.

*4. The Many-Minds Interpretation of Quantum Mechanics*

In this section I want to explain how the philosophy of quantum mechanics gives us reason to adopt a view of reality which makes the rejection of (A) and (B) quite natural.

Within quantum mechanics any physical system, such as a moving electron, is characterized by a mathematical device called a state vector, or wave function. This function does not specify exact values for the position or velocity of the electron. Instead it specifies the probabilities that the electron will turn up with any of a number of different positions or velocities when the quantities are measured.

Quantum mechanics also contains an equation, called Schrödinger's equation, which specifies how the wave function of the electron will evolve smoothly and deterministically over time. This is analogous to the way Newton's laws of motion determine the evolution of a body's posi-

tion and velocity over time. Except that, where Newton's laws deal with actual positions and velocities, the Schrödinger equation describes the evolution of *probabilities* of positions and velocities.

So quantum mechanics, as normally understood, needs to appeal to another kind of process, in order to turn probabilities into actualities. This second process is commonly known as the "collapse of the wave function", and is supposed to occur when a measurement is made. So, for example, if the electron collides with a sensitive plate, and registers in a particular position, the probability for that position instantaneously jumps to one, and for all other positions to zero.

However, if you stop to think about it, this scarcely makes sense. What qualifies the collision with the plate as a "measurement"? After all, the joint system of plate plus electron can itself be viewed as a large collection of microscopic particles. And as such the joint system will be characterized by a probabilistic wave function, which will then evolve smoothly in accord with Schrödinger's equation. From this perspective, there will then be no collapse into an actual position after all, but simply probabilities of the electron being in different places on the plate once more.

In practice, most physicists assume that a wave-collapsing measurement occurs whenever a big enough physical system is involved. But how big is big enough? It seems arbitrary to draw the line at any particular point. And even if we did know where to draw it, we wouldn't have any principled physical explanation of why it should be drawn there.

This is the moral of "Schrödinger's cat". Imagine that some unfortunate cat is put in a chamber which will fill with poison if the electron registers on the left half of the plate, but not if the electron registers on the right half. Until the wave function collapses, reality remains undecided between the two possibilities, alive or dead. So when does reality decide? When the electron hits the plate? When the poison kills the cat? Or only when a human enters the room and sees if the cat is alive or dead? Nothing in quantum mechanics seems to tell us when or why the collapse occurs.

Some philosophers hold that quantum mechanics is incomplete, and that in addition to the quantities recognized by quantum mechanics there are various further "hidden variables". Hidden variables can avoid the problem of Schrödinger's cat, by implying that it is fixed from the start whether the cat will be alive or dead. However any hidden variable theory

which is consistent with the experimental data (in particular with non-local correlations) will be in tension with special relativity, since it will require the transmission of influences across space-like intervals.

There is another way. Suppose we take quantum mechanics to be complete, but deny that the wave function ever collapses. That is, reality never does decide between the live and dead options for Schrödinger's cat. The electron keeps positive probabilities both of being on the left and of being on the right of the plate, the cat keeps positive probabilities both of being alive and of being dead, and your brain keeps positive probabilities both of seeing the cat alive and seeing it dead.

At first sight this might seem to contradict our experience. When we look, we either see a live cat or a dead cat, not some combination of both. But we need to ask: what exactly *would* it be like to have a brain whose wave function evolved into a superposition of seeing a live cat and seeing a dead cat? There is no obvious reason to suppose that it would involve some kind of fuzzy experience, like seeing a superimposed photo of a live and dead cat. Instead, perhaps it would be like being two people, one of whom sees a dead cat, and the other a live cat.

This is the "many minds" theory of quantum mechanics. According to this theory, when an intelligent being interacts with a complex quantum system, its brain acquires a corresponding complexity, each element of which then underpins a separate centre of consciousness. One aspect of your brain sees a live cat, another a dead cat.

Of course, if these two consciousness are both present in reality, we need some account of why there is no direct evidence for this. But the many-minds theory can explain this. There are possible experimental circumstances which would demonstrate that your brain is in a superposition of both live cat and dead cat perceptions. But with a system as complex as a human brain, these experiments are far too difficult to carry out. And this is why there is no direct evidence for the duality of your brain. Even though both elements are present in reality, it is too hard to arrange the precise experimental circumstances which would allow this to manifest itself.

The mathematical underpinnings of the many-minds theory were laid out by Hugh Everett (1957) nearly forty years ago. Everett's ideas have been characterized by a number of writers as a "many worlds" theory. But this is not the best way to read Everett's suggestion. The idea that the world splits in quantum measurements creates as many problems as it

solves. Apart from anything else, it still ascribes a special status to measurements. A better thought is that there is just one world, characterized by an evolving wave function, in which the only things that split are the conscious experiences of the brains involved in that wave function.

It is important to realize that there is nothing inherently dualistic about this many-minds theory. If you think that the minds are physical systems, as I do, then you can carry on thinking this on the Everett view of reality. The resulting position will maintain that reality is exhausted by the evolving wave function of the universe, but will point out in addition (as does conventional physicalism about the mind) that certain subsystems of this physical reality have the kind of complexity that constitutes mental experience.

Where the Everett view differs from conventional thought is simply in holding that there are many more physical systems with this kind of complexity than is normally supposed. When your brain interacts with a chancy set-up, like Schrödinger's cat, for example, it evolves into a number of consciousness-underpinning physical systems, rather than just one. The later systems will share memories with the earlier system (of walking over to the cat-box, opening it, etc.), but will have divergent memories thereafter (seeing it alive versus seeing it dead). An amoeba splitting into two provides a simple partial analogy. And in mental terms we can think of the earlier self as having a number of successors, none of whom is identical with the earlier self, but all of which are connected to it. (Derek Parfit's account of personal survival fits very well with the many-minds view. See Parfit, 1984.)

So reality contains a set of branching perspectives. In reality, all outcomes occur in any chancy situation. But a given conscious perspective only ever registers one such outcome. For any continuing conscious individual, therefore, the world seems to evolve normally. Cats turn out to be either alive or dead, coins land either heads or tails, and so on. But from a God's-eye point of view all the outcomes occur, since all the different perspectives are equally present in reality. Note that the entities that branch over time, or "split", are systems-with-a-certain-perspective-on-reality, not reality itself, which can be seen as the evolving sum of those perspectives.

These remarks only scratch the surface of the many-minds interpretation. I have been largely following Michael Lockwood's elaboration of Everett's ideas (Lockwood, 1989). A variant many-minds theory has been

developed by David Albert and Barry Loewer (1988, 1991); their version, however, requires a commitment to dualism, for what seem to me bad reasons (see Papineau, 1995). For further discussion of the many-minds interpretations approach, and of the differences between the Lockwood and Albert-Loewer approaches, see Lockwood *et al.* (1995).

*5. Probabilities and Many Minds*

I don't know how seriously we should take the many-minds theory. My aim in this paper is only to establish the hypothetical claim: *if* you accept the many-minds view, *then* the puzzles raised in Sections 2 and 3 go away.

First, however, I need to say something general about the many-minds view and probabilities. Consider the example we started with. You are faced with a quantum mechanical device, with a chance of 0.8 for H and 0.2 for T. You can bet either way, at evens. So you back H.

On the many-minds view, once you discover the result (i.e., you interact with the quantum-mechanical device after it has been "spun") you split into two successors, one who sees H and wins £1, and one who sees T and loses £1. The former outcome has a probability of 0.8, the latter a probability of 0.2.

The initial issue here is whether the many-minds theories is still entitled to view these numbers as *probabilities*. It is true that the quantum-mechanical formalism still imposes a 0.8–0.2 measure on the two outcomes. But in what sense can this be viewed as a *probability* measure, if both outcomes are sure to occur in reality? Isn't probability a measure of which outcome is most likely to *win* the competition to become real? And doesn't the many-minds view deny that any outcome wins in this sense?

I have addressed this issue in another paper (Papineau, 1995). I argue there that the many-minds view is no less entitled than conventional thought to regard the relevant numbers as probabilities. Probability is a rather puzzling notion within the many-minds view. But it is just as puzzling, and for just the same reasons, within conventional thinking.

It is true that, according to the many-minds views, all alternative outcomes with non-zero probability will occur, whereas conventional thought says that just one will. But I argue that, while this is a real difference (indeed it is a restatement of the difference between the many-minds

and conventional views), it is not a difference that matters to anything else we do or think with probability. So it is a weak reason for denying the many-minds theory the notion of probability: it does nothing to show that many-minds theory cannot treat the relevant measure as a probability in every other respect.

To see this, note that probability enters into our reasoning in two ways. (1) We infer probabilities from observed frequencies. (2) We use probabilities to inform our choices.

Let me take these in turn. The logic by which we infer probabilities from observed frequencies is not well understood. In practice, we judge that the probability is close to the observed frequency, and hope that we are not the victim of an unlucky sample. The many-minds theorist can recommend that any thinker should do just the same. As to the justification for this strategy, the many-minds theory can observe that conventional thought offers no agreed rationale. Moreover, the alternatives on offer (Fisherian, Neyman-Pearsonian, Bayesian) are equally available within the many-minds approach.

Now take the relevance of probability to choice. Consider our original example once more. You might doubt whether the many-minds theory can explain why you should bet on H. If you are inevitably going to have a successor whose coin shows T and so loses £1, alongside the successor who sees H and so wins £1, then what's so good about betting on H? Wouldn't you achieve just the same pair of results (namely, both winning and losing £1) by betting on T?

But the many-minds theorist can simply say that it is a primitive fact about rational choice that the best action is the one that maximizes the probabilistically expected gain, that is, the weighted average of gains weighted by the chance of each outcome. Since H has a chance of 0.8 and T only 0.2, the expected gain of backing H is greater than that of backing T, which is why we should bet on H.

You might feel inclined to object to the many-minds theory plucking this "primitive fact about rational choice" from thin air. But before you do, note that this primitive fact is nothing but principle (B) from Section 2 once more. And remember that, however we turned it in Section 2, we couldn't find any justification for principle (B) within conventional thought either. Along with Peirce and Putnam, we were forced to accept it as a "primitive fact about rational choice". It would scarcely be reasonable

to deny this primitive fact to the many-minds theory, while allowing it to conventional thought.

*6. Puzzles Dissolved*

I have just pointed out that the many-minds theory is *no worse off* than conventional thought at explaining uncertain choice. I now want to show how the many-minds theory is *better off* than conventional thought on this issue.

In Section 2, I argued that the really puzzling thing about principle (B) isn't just that it has to be taken as primitive. Rather it is that it is in tension with the further principle (A), according to which the best action *isn't* the one that will *probably* gain, but the one that *actually* will. This tension led me to observe that we would be better off if we could jettison (A), and explain all our choices by (B) alone.

In Section 3, I made the same points about the knowledge-relative principle (C) in relation to principle (B). Principle (C) is in tension with (B), since (B) says the best action maximizes single-case expectation, not knowledge-relative expectation. So I suggested that perhaps we should jettison (B) along with (A), and explain everything by (C) alone.

At that stage it seemed to make little sense to jettison (B) and (A). Surely it is advisable to bet with the single-case odds, and even better to get what you want. However, from the many-minds point of view it makes perfect sense to jettison (B) and (A).

My simple initial example illustrated the tension between (A) and (B). Given an 0.8 chance of H, then betting on H at evens is the $\text{best}_B$ option. But this option isn't $\text{best}_A$ (it doesn't actually gain anything) if you get outcome T.

However, note that the whole notion of the $\text{best}_A$ option, and hence the resulting tension, is premised on the assumption that the device will *either* display H, *or* T, but not both. After all, the $\text{best}_A$ option is H *if* the device displays H, and T *if* it displays T. On the many-minds view, however, the device displays both results: H relative to one of your successors, and T relative to the other. So the notion of the $\text{best}_A$ option disappears, and we are left with the single idea that the best option is simply the option which maximizes your gain over the two outcomes, weighing them by their chances—that is, the $\text{best}_B$ option. There is no further worry that this won't really be best, if you get the unlucky outcome

rather than the more probable one, for we are no longer supposing that you will get just one of these.

Similarly with the more complicated example which illustrated the tension between (B) and (C). Given a 0.6 probability that your device is H-biased, you prefer the $best_C$ option of backing H. But what if your device is T-biased? Then backing H won't be $best_B$.

Again, this worry is premised on the thought that the device is either H-biased or T-biased but not both. According to the many-minds point of view, however, your device is in a superposition of both types of bias, since its quantum-mechanical manufacture gave a positive probability to both. So the notion of the $best_B$ option disappears, and the only sense left in which a choice can be best is the $best_C$ sense of maximizing gain over the two types of bias, weighed by their respective knowledge-relative probabilities. The thought that this choice won't really be best, if the actual bias is the less probable one, dissolves along with the supposition that the device definitely has one bias rather than the other.

*7. The Source of the Puzzles*

On the conventional view of the world, the single-case probabilities (the chances) relevant to any choice themselves evolve over time. These probabilities are naturally thought of as the probabilities fixed at any time by all the facts that are determinate at that time. Consider once more the outcome H on a "spin" from a device from the machine, to be performed at noon, say. Before the machine manufactures the device, and the device's bias is not yet fixed, the chance of H is 0.66 ($0.6 \times 0.9 + 0.4 \times 0.3$). Once the device has been made, with an H-bias or a T-bias, the chance of H is either 0.9 or 0.3. And just after noon, and from then on, the chance of H is either 0 or 1.

In practice, however, any agent acts on the basis of his or her current knowledge-relative probabilities. These need not be (though they may be) equal to the putative single-case chances at the time of the decision or at any later time. This potential divergence is the source of the puzzles we faced earlier. We feel that, if the current or later single-case probabilities are different from the agent's knowledge-relative probabilities, then the agent's deliberations will be based on less than ideal information and may well lead to the wrong option. And this then makes it difficult to explain the worth of knowledge-relative choices.

From the many-minds perspective, however, the single-case probabilities relevant to a choice will never diverge from the knowledge-relative probabilities. This is because these single-case probabilities do not evolve over time. All that evolves, on the many-minds view, are your opportunities to interact with chancy set-ups, and thereby to split into successors who acquire different information.

So, for example, once the device has been manufactured, you are in principle able to turn it over and see whether it reads "h-series" or "t-series". Conventional thinking assumes that if you do this (i) you will either read "h-series" or "t-series", but not both, and (ii) your turning it over won't affect the bias; it therefore concludes that the device is *already* one or the other, and the single-case probability of H is therefore already 0.9 or 0.3. From the many-minds point of view, however, *both* biases will be observed if you turn the device over, each by a different successor of yours, so there is no reason to think that, whether or not you actually turn it over, H now has any probability different from your knowledge-probability of 0.66 (i.e., 0.9 $\times$ the 0.6 probability of H-bias, plus 0.3 $\times$ the 0.4 probability of T-bias).

The same applies even when the device is "spun" at noon and you observe the result. At noon you will be able to observe either H or T. So conventional thought concludes that at noon the chance of H must be either 0 or 1. But the many-minds theory says that if you observe the result at noon you will then have two successors (or four, if you turned the device over to see the bias first), and (either way) the sums will still show that at noon H retains its original knowledge-relative probability of 0.66.

This doesn't imply that it wouldn't be worth finding out the bias, or even the actual result, before betting, if anybody were generous enough to leave the bet open that long. The many-minds view agrees with conventional thought that, other things being equal, it is better to get more information before acting. But from the many-minds point of view the advantage of waiting-and-seeing isn't that you will then be able to bet in line with the *uniquely real* single-case odds (0.9 or 0.3, 1 or 0), as opposed to the *various possible* single-case odds your limited information doesn't decide between. Rather it is that you will be better able to maximize your expected gain *over all the possibilities* if you can delay acting until after you "branch", and thereby allow your successors each to choose the option that will maximize the expected gain on their branch.

So, for example, if you can turn the device over before betting, your "H-bias successor" will choose H, with expected gain of £0.80, and your

"T-bias successor" will choose T, with expected gain of £0.40, and the weighted average over your two successors is thus £0.64, rather than the £0.32 offered by backing H immediately. The point is that your successors will do better overall if they are allowed to choose different options according to their individual circumstances, rather than your picking one option on behalf of all.

This is just the Ramsey-Good result again. But note how it now makes much more sense than it did in Section 3. At that stage it seemed odd to *justify* waiting-and-seeing in terms of its maximizing *knowledge-relative* expectation. (Isn't waiting-and-seeing better simply because it allows you to choose according to whichever of the possible single-case probabilities is *real*?) But from the many-minds point of view all the possibilities are real, and the point of waiting-and-seeing is indeed that it maximizes expected gain across all possibilities weighted by their knowledge-relative probabilities.

This now makes it clear why the many-minds perspective has no need, indeed no room, for any standard for assessing choices apart from principle (C). Any agent will do best for itself and its successors if it acts so as to maximize probabilistically expected gain across all the possibilities in its future with non-zero knowledge-relative probability—that is, if it conforms to principle (C). The only sense in which limited knowledge can be bad is that sometimes the best action is to *defer action*. Agents sometimes do $\text{best}_C$ by first "splitting", and then allowing their successors to choose individually—but then these successors simply choose according to principle (C) in their turn.

When agents do "split" in this way, it is true their successors will end up with different knowledge-relative probabilities for the relevant outcomes. When you don't know the bias of your device, your knowledge-relative probability for H is 0.66. Once you turn it over, your "h-series" successor will have a knowledge-relative probability of 0.9 for this same outcome, while your "t-series" successor will have a knowledge-relative probability of 0.3. But these chances, unlike the putative evolution of single-case probabilities within conventional thought, do not devalue the knowledge-relative choices made by the original "pre-split" agents. In the overall future of your pre-split self, the probability of H-at-noon does not *alter* from 0.66 to 0.9 (or to 0.3, as the case may be). Rather it *factors* into your two successor branches, according to the sum: $0.66 = 0.9 \times$ the 0.6 probability of H-bias, plus $0.3 \times$ the 0.4 probability of T-bias. More generally, if R is the outcome of interest, and $K_i$ ($i = 1, \ldots n$) are the possible results of an agent's

acquiring more information, then the agent's original probability for R, Pr(R), will factor according to the equation $P(R) = \Sigma_i P(R/K_i)P(K_i)$, when its successors acquire the information in question.

In this paper I have tried to show how the many-minds theory makes better sense of uncertain choice than conventional thought. Let me conclude by briefly noting that my remarks also point to two further ways in which the many-minds theory promises a philosophical advantage. First, it offers an obvious explanation for probabilistic *conditionalization*, in that it simply falls out of the metaphysics and our basic principle (C) that agents ought to set their new probabilities for R equal their old $Pr(R/K_i)$ when they discover $K_i$. Conventional approaches to probability and decision have difficulty accounting for conditionalization. Second, the many-minds view implies that the overall objective probability of any given result *does not change over time*, since all "branches" exist and the weighted sum of any Pr(R) over all branches remains constant. It is a demerit in conventional thought that it takes a fundamental physical quantity like objective probability to evolve asymmetrically over time. However, these are topics for two further papers.[4]

*David Papineau*

*King's College*
*London*

NOTES

1. The expected gain of an action is the weighted average of the gain that the action will yield in each alternative possibility (such as H or T), weighted by the chances of those possibilities. In this paper I shall stick to examples in which the chances of the alternative possibilities do not depend on the agent's choice, so as to by-pass the debate between causal and evidential decision theory.

2. Although it is convenient to introduce knowledge-relative probabilities with talk of how "often" given results tend to occur in given kinds of situations, I do not endorse any kind of frequentist reduction of these probabilities. I think the frequency theory of probability is hopeless, for familiar reasons (cf. Papineau, 1995, p. 244).

3. By "single-case expected gain" I simply mean "probabilistically expected gain" defined in terms of chances as in n. 1 (above); "knowledge-relative expected gain" can then be defined similarly, but with the alternative possibilities weighted by their knowledge-relative probabilities, rather than by their chances.

4. I would like to thank Helen Beebee, Scott Sturgeon and the members of the King's College London Philosophy Department for help with the ideas in this article.

## REFERENCES

Albert, D. and Loewer, B. (1988); "Interpreting the Many Worlds Interpretation", *Synthese*, **77**, 195–213.

Albert, D. and Loewer, B. (1991): "The Measurement Problem: Some 'Solutions'", *Synthese*, **86**, 87–98.

Everett, H. (1957): "'Relative State' Formulation of Quantum Mechanics", *Review of Modern Physics*, **29**, 454–62.

Good, I. J. (1967): "On the Principle of Total Evidence", *British Journal for the Philosophy of Science*, **18**, 319–21.

Lockwood, M. (1989): *Mind, Brain and Quantum* (Oxford: Basil Blackwell).

Lockwood, M., with Brown, H., Butterfield, J., Deutsch, D., Loewer, B., Papineau, D., and Saunders, S. (1996): "Symposium on 'Many-Minds' Interpretations of Quantum Mechanics", *British Journal for the Philosophy of Science*, **47**.

Parfit, D. (1984): *Reasons and Persons*, Oxford: Oxford University Press.

Papineau, D. (1995): "Probabilities and the Many-Minds Interpretation of Quantum Mechanics", *Analysis*, **55, 239–46**.

Peirce, C. S. (1924): "The Doctrine of Chances" in M. R. Cohen (ed.) *Chance, Love and Logic*, New York: Harcourt.

Putnam, H. (1987): *The Many Faces of Realism*, La Salle, Open Court.

Ramsey, F. P. (1990): "Weight or the Value of Knowledge", *British Journal for the Philosophy of Science*, **41**, 1–4.