Why you don't want to get in the box with Schrödinger's cat

David Papineau

1. Introduction. In his 'What is it like to be Schrödinger's cat?' (2000), Peter J. Lewis argues that the many minds interpretation of quantum mechanics has absurd implications for agents facing chancy life-or-death decisions.

By way of an example, Lewis imagines your being invited to join Schrödinger's cat in its box for an hour. This box will either fill up with deadly poison fumes or not, depending on whether or not some radioactive atom decays, the probability of decay within an hour being 50%. The invitation is accompanied with some further incentive to comply (Lewis sets it up so there is a significant chance of some pretty bad but not life-threatening punishment if you don't get in the box). Lewis argues that the many minds theory implies that you should get in the box with the cat, despite this making it 50% likely you will die.

His reasoning is as follows. In general, the many minds theory holds that a conscious being who observes the outcome of a chancy either-or experiment will evolve into two successors (in different 'branches' of reality), each of whom observes one of the possible outcomes. Moreover, the theory advises you to favour choices in such chancy situations in proportion to the probability that they will bring good or bad results to your various successors. But in a life-or-death case like getting into the box, you will only have one successor, since one of the outcomes (the atom decays) will ensure your death. So it seems that the many minds interpretation advises you to get in the box with the cat, since it is certain that your only successor will emerge unharmed.[1[1]]

My aim in this paper is to show that the many minds interpretation does not in have this counter-intuitive death-defying implication. It is certainly true that the many minds interpretation of quantum mechanics violates many of our most basic assumptions about reality. But it does not go so far as to urge us to risk death with abandon.[2[2]]

1[1] This line of thought has recently been endorsed by another Lewis. In his Jack Smart Lecture in Canberra in June 2001, entitled 'How Many Lives Has Schrödinger's Cat', David Lewis argues similarly that no-collapse accounts of quantum mechanics imply that agents should never expect death, and so need not take death risks into account in their decision-making.

2[2] Would it damn the many minds theory if it did have the death-defying implication? It is not clear. David Lewis regards the implication as a reason to wish the theory false, but points out that this is not the same as a reason to believe it false. Peter J. Lewis, at the end of his paper, suggests that the implication means the theory cannot be publicly tested, but his argument for this strikes me as less than conclusive. I myself think the implication would weigh against the theory simply because it runs so counter to our evolved decision-making inclinations: on the presumption that these inclinations have evolved because they served our ancestors' interests, it surely counts against a theory if it implies that those interests would have been much better served by quite different inclinations. However, there is no need for me to nail down this

2. Why Defy Death? Some commentators hold that the many minds theory does not so much as allow talk of probabilities for different outcomes in chancy situations (and thus a fortiori cannot maintain the principle that rational agents should weigh choices in proportion to the probabilities of good and bad results). The worry here is that probabilities are by definition measures of the potential for given outcomes to become actual rather than non-actual; but on the many minds theory all physically possible outcomes will inevitably become actual (in different 'branches' of reality); which makes it hard to see how the many minds theory can give a probability different from unity to any physically possible outcome. (Cf. Albert and Loewer, 1988.)

This is not Peter J. Lewis's worry. Citing Papineau (1995) in support, he observes that it would be no more puzzling that agents should weigh their choices in proportion to the 'probabilities' attached to outcomes by the many minds theory, than that they should weigh their choices in proportion to probabilities as conceived within conventional metaphysics[3[3]]. If you are betting simply on whether the cat will live, and don't have to get in the box, then you should bet just in case the odds are better than evens, says Peter J. Lewis, for the standard reason that so betting will produce a net expected gain, averaging over your successors, weighted by their 'probabilities'.[4[4]]

So Peter J. Lewis's worry pertains specifically to choices where one of the physically possible outcomes is your death. It is specifically these outcomes that he thinks the many minds theory requires us to ignore.

---

issue here. Since I deny that the many minds theory has the death-defying implication to start with, I don't need to decide how far it would damn the theory if it did.

[3[3]] He also attributes to me the thought that the many minds theory makes risky choices just as puzzling as they are given conventional metaphysics, as well as no more puzzling. However, I actually think that the many minds theory makes risky choices much less puzzling than conventional metaphysics does. This point is only mentioned in passing in Papineau (1995)—see footnote 5—but it is elaborated at length in Papineau (1997).

[4[4]] Here Peter J. Lewis concedes more to the many minds theory than the other Lewis does. David Lewis regards the principle that you should match your expectations to the quantum mechanical 'intensities' (his term) as an addition to the more familiar principle that you should match your expectations to chances, and then observes that no-collapse theorists cannot justify their new 'intensity' principle in terms of the old chance principle, because of the radical differences between their conception of intensity and the familiar notion of chance. I find this a puzzling complaint. Any no-collapse theorist will surely regard the 'intensity' principle as a replacement for the old chance principle, not as an addition to it. Perhaps the metaphysical differences between their 'intensities' and the more familiar chances should make us regard this as a matter of elimination rather than reduction, but either way no-collapse theorists need only uphold one principle, the basic 'intensity' principle. Moreover, they can then point out that its lack of justification exactly parallels the chance principle's lack of justification within conventional thinking.

Why does he think this? It is not entirely clear. The crucial premise, as I stated it above, is his assumption that:

(I) the many minds theory implies you should favour choices in chancy situations in proportion to the probability that they will bring good or bad results to your various successors.

However, there is an obvious alternative premise, on which the death-defying implication does not follow:

(II) the many minds theory advises you to favour choices in chancy situations in proportion to the probability that they will produce good or bad outcomes whether or not you have a successor to experience those outcomes.

If we assume (II) rather than (I), then the many minds theory will no longer imply that you should get in the box with Schrödinger's cat. Instead it will urge you to reason as follows: if I get in the box, there are two physically possible outcomes, each with 50% probability—I emerge unharmed, and I die; the first is OK, but the second is very very bad; so the expected outcome of getting in the box, on weighted average across the two outcomes, is still very bad; so I should turn down the invitation to get in the box (even given the risk of the pretty bad but not life-threatening punishment for refusing the invitation).

This reasoning exactly matches that of conventional thinking (even if the underlying metaphysics is different). More generally, on assumption (II) the many minds theory will deliver exactly the same advice about choices as conventional thinking, even in life-or-death situations.

So this issue hinges entirely on Peter J. Lewis's preference for assumption (I) over (II). I can think of three possible rationales for this preference. Let me consider them in turn.

3. Hedonistic Values. Perhaps the only things of value are experiences; because of this, outcomes in which you do not undergo any experiences are of no value, positive or negative, and so should be eliminated from expected utility calculations, given the many minds theory.

That this may be Peter J. Lewis's rationale for ignoring fatal outcomes is suggested by his taking care to specify that in his example you do not care about the suffering of any (successors of) friends and relatives who may mourn your death (p. 26). This specification would make sense if experiences were the only items that can imbue future outcomes with value. In that case, sufficiently altruistic agents could care about outcomes in which others suffer, even if they themselves are dead; but no rational agents could care about outcomes in which nobody suffers.

I am not sure whether this hedonistic perspective lies behind Peter J. Lewis's thinking.[5][5] But in any case the hedonistic perspective is clearly mistaken. It is

surely cogent, even if slightly unusual, for me to care, say, about the continued flourishing of my secret garden after I die, even if I know that no conscious being will ever enjoy it.

Nor is it even necessary to consider such recherché cases. If one outcome is valuable because it contains my future experiences, surely an alternative outcome which lacks those experiences is of lesser value, simply by comparison with the first outcome. Since expected utility calculations hinge on relative utility values rather than absolute ones, I should be concerned about death as long as the outcome where I die is given less utility than the one where I survive, whatever the absolute value.

Apart from the above points, the hedonistic rationale for supposing that the many-minds theory ignores death-risks is open to the objection that it seems to apply equally well to conventional thinking. If experienceless outcomes are of no consequence, positive or negative, then shouldn't conventional thinking also ignore them? Thus, on conventional thinking: either I will emerge from the box unharmed, or I will die (but not both); the first outcome would be OK, while the second would be of no consequence; so I should get in the box, since there no downside to weigh against the 50% chance I will get a good outcome. Since conventional thinking does not, I take it, actually have this death-defying implication, it follows that there must be something wrong with the hedonistic rationale.

4. Premature Renormalization. The second possible rationale for (I) focuses on the probabilities attaching to fatal outcomes, rather than to the utilities. On p. 26 Peter J. Lewis says that, when you are invited to get in the box with the cat, you should attach a probability of 100% to emerging unharmed. He then continues:

'Before I said that there is a 50% chance of each outcome; now I am saying there is a 100% chance of surviving unharmed. But there is no real conflict here, since the probabilities we are talking about here are subjective probabilities relative to a particular observer. If an observer has successors who see each potential outcome of an experiment, then the usual quantum mechanical probabilities are the ones which are relevant for that observer. However, if the observer only has successors for some of the potential outcomes, then the relevant probability measure covers only those outcomes; the usual quantum mechanical probabilities for those outcomes need to be renormalized until they sum to one.'

I see no reason whatsoever to view the probabilities in this way. True, rational agents act on the basis of subjective probabilities. But ideally their subjective probabilities will match objective ones, as Peter J. Lewis himself allows. The real question is thus: which objective probabilities should ideally constrain subjective

---

5[5] There are indications that David Lewis is thinking in this hedonistic way: throughout his lecture he insists, without further argument, that the rule connecting expectations with 'intensities' must be viewed as a rule governing expectations of experience, not expectations of what will happen whether experienced or not (cf. especially his footnote 23).

probabilities? And the natural answer is: the objective probabilities relative to the agent's branch of reality at the time when the agent makes his or her decision.[6[6]]

However, if we accept this answer, then it follows that, when you are invited to get in the box with the cat, you should adopt a 50% subjective probability for emerging unharmed, rather than 100%, since prior to your action that is your objective probability of survival relative to your branch of reality.

Of course, after you get in the box, and in due course have an unharmed successor, then that successor will come to attach a 100% subjective probability to survival. For that successor will now be on a more specific branch of reality that you originally were prior to your action, and surviving the box has a 100% objective probability relative to this successor branch.

However, there is no obvious reason to infer, from the fact that your sole surviving successor should later on attach a 100% subjective probability to survival, that you yourself should attach that same subjective probability to survival before getting in the box, when there is still a 50% objective probability of your death relative to your branch of reality.

5. Too Few Worlds. So far I have been assuming that, if I get in the box, there is a future branch of reality on which I die, even though I have no experiences there. Perhaps this assumption can be questioned. And then I won't have a reason to stay out of the box. If the only future branches of reality are ones where I am alive and experiencing, then my utility calculations need not consider physical possibilities where I am not alive, for there are none, and assumption (II) above collapses into assumption (I).

Why might the very existence of a future branch of reality depend on my being alive to experience it? Well, one possible motivation for a 'many minds' approach to quantum mechanics is the worry that the future will not factor naturally into 'branches' without the help of conscious minds. According to this line of thought, multi-faceted futures do not automatically configure into unique sets of 'branches', since there are many possible proper ways of decomposing those futures into elements, most of which correspond to nothing humanly recognizable; we get discernible 'branches' only in virtue of minds that experience specific elements of those futures as definite. In technical terms, the idea is that minds are needed to pick out some 'preferred basis' with respect to which the future decomposes into recognizable scenarios. (Cf. Barrett, 1999, ch. 7.) More intuitively, you might think of a future experiencing mind, on this motivation, as akin to a spotlight which illuminates some part of the overall future as a definite 'branch'.

On this motivation for 'many minds' talk, then, there will arguably be no branches where there are no minds. And then, as I said, you would indeed have no

---

6[6] For further discussion of this issue, see Beebee and Papineau (1997), Papineau (1997).

reason to stay out of the box, since there is no future branch on which you die, for lack of any experiencing mind to constitute it as a branch.[7[7]]

However, it does not seem that this is Peter J. Lewis's reason for ascribing the death-defying implication to the many minds theory. For he says (footnote 1) that the argument of his paper applies, not just to interpretations of quantum mechanics which explicitly advertise themselves as 'many minds' theories, but also to the 'many worlds' interpretation associated with Bryce DeWitt, and to the 'many histories' interpretation associated with Murray Gell-Mann, James Hartle, and W.H. Zurek. None of these further interpretations supposes that the future is only constituted into alternative branches when there are minds around to discern those branches. Rather, they appeal to some more objective grounding for the 'preferred basis' needed to factorise the future into recognisable scenarios, such as the possibility of 'decoherent histories'.

Moreover, even those views which do go under the banner of 'many minds' do not necessarily suppose that future branches are constituted by experiencing minds. There are various independent rationales for adopting the terminology of 'many minds', apart from the promise of a mind-dependent solution to the 'preferred basis' problem.[8[8]] Given this, there is no reason why 'many minds' views should not maintain that future branches are constituted by the fabric of non-mental reality itself, in such a way that branches can exist even where there are no minds. And then the death-defying implication will be avoided once more. If experienceless branches can exist, then your getting in the cat's box will ensure that there will be a branch of reality, with 50% probability, in which you will shortly die, which gives you every reason to stay out of the box.

6. Conclusion. Whichever way we turn it, Peter J. Lewis seems to have no good reason for ascribing the death-defying implication to the many minds theory. This ascription would only be justified if the many minds theory assumed that future branches are constituted by experiencing minds. However, the many minds theory is not committed to this assumption, nor does Peter J. Lewis presume that it is.

[7[7]] Couldn't other people's experiences constitute this branch (whether or not you are concerned about the distress those people will suffer—cf. section 3 above)? Yes, but the difficulty still remains for 'branches' on which no beings at all have experiences. It wouldn't be very reassuring that the many minds theory tells you not to get into Schrödinger's box, if at the same time it sees no danger in actions that will probably destroy all sentient life.

[8[8]] One such rationale is to distance the theory from talk of 'world-splitting', and make it clear that 'branchings' of reality are always an entirely local matter. The whole universe doesn't split into two when a detector interacts with a particle in a superposition: all that happens is that the detector gets into a superposition too. And similarly when a conscious observer interacts with the detector: we don't get two overall universes, but simply one universe in which that conscious observer gets into a superposition.

King's College London, Strand, London
WC2R 2LS, UK
david.papineau@kcl.ac.uk

References

Albert, D. and Loewer, B. 1988. Interpreting the many worlds interpretation. Synthese 77: 195-213.

Beebee, H. and Papineau, D. 1997. Probability as a guide to life. Journal of Philosophy, 94: 217-43.

Lewis, Peter J. 2000. What is it like to be Schrödinger's cat? Analysis, 60.1: 22-9

Lewis, David. 2004. How many lives has Schrödinger's cat? Jack Smart Lecture, Canberra, June 2001. To be published in March 2004 in Australasian Journal of Philosophy, 82.

Papineau, D. 1995. Probabilities and the many minds interpretation of quantum mechanics. Analysis, 55.4: 239-46.

Papineau, D. 1997. Uncertain decisions and the many-worlds interpretation of quantum mechanics. The Monist, 80: 97-117