Saunders and Wallace on Everett and Lewis

Paul Tappenden[1]

18 June 2007

Simon Saunders and David Wallace are attempting to use a modified form of David Lewis's analysis of personal fission to support the thought that prior to undergoing Everett branching an informed subject can be uncertain about which outcome s/he will observe. I argue that the seductive appeal of the idea is an illusion arising out of failure to distinguish between the provenance and reference of the relevant observers' utterances.

Simon Saunders and David Wallace suggest that a subject about to knowingly undergo branching in the Everett multiverse can be understood to be uncertain about what s/he will experience if David Lewis's view of the transtemporal identity of persons through fission is modified (Saunders and Wallace 2007, Wallace 2005a: sec. 3.4, 2005b:14; Lewis 1976). Wallace has what can appear to be an independent argument for pre-measurement uncertainty in making intelligible Hugh Everett III's 'relative state' interpretation of quantum mechanics ; it is an argument from linguistic interpretative charity. I shall not be challenging that argument here but I shall have something to say about it in relation to the Lewis-based idea. Wallace's 2005 papers only make brief mention of this idea but it is the focus of attention of Saunders' and Wallace's (2007).

Lewis used the world-tube (or 'worm') version of transtemporal identity which takes a persisting object to be an aggregate of temporal parts or 'stages'. For personal identity, the cement which holds the aggregate of person-stages together is an 'R-relation' between temporally juxtaposed stages. In a case of genuine personal fission some temporal stage S at time T has multiple successors at a later time T+ which are all R-related to S. Independently of any concern about Everett's interpretation of quantum mechanics there has been discussion of the possibility of such personal fission in imaginary cases of partial brain transplants and malfunctioning teleporters, dealt with at length by Derek Parfit (1984). Lewis argued that each of these multiple successors can be understood to be a stage of a distinct persisting person who has their origin prior to T and who persists to at least T+. The histories of these distinct persons thus overlap prior to T+. The stage S is a stage of many persons, one for every downstream branch of the 'world-tree' of which the pre-fission segment is the trunk.

[1] Middlesex University, London, UK paulpagetappenden@googlemail.com

Putting this idea into the context of a quantum measurement in Everett's multiverse where a measurement of the spin of a particle relative to some arbitrarily given spatial direction is made, there are as many observers as there are downstream branches. In realistic Everettian terms that means that there will be a vast and indeterminate number of observers. However, leaving aside very low amplitude branches, the observers will partition post-measurement into a set of observers seeing spin-up and and a set of observers seeing spin-down. It will do no harm to simplify this setup for the sake of clarity. So imagine the idealised situation where there is a single fission into two branches with no subsequent branching. What is important here is tracking identity so we can set aside the quantum-mechanical amplitude.

Our observer, Hydra, is assumed to believe the Everett interpretation and to be fully informed about the relevant aspects of her quantum-mechanical predicament. She has prepared her Stern-Gerlach apparatus at time T for a measurement of the residual x-spin of a silver atom and the result is going to be, according to the Everett interpretation, that at the later time T+ Hydra will have two 'successors' one of whom will see the result UP and the other the result DOWN. On the Lewisian analysis of fission there are in this scenario two observers who we can label HydraUP and HydraDOWN. Following Lewis, these Hydras have distinct stages at T+, where the outcomes UP and DOWN are respectively seen, and common stages up to and including T. Saunders' and Wallace's claim is that at T each of the Hydras can truly say 'I am either HydraUP or HydraDOWN but I do not know which '. Thus each Hydra at T is subject to ignorance about which person she is and this justifies the assertion that each is uncertain about what she is going to see.

The idea can be seductive but we need to look closely at how people are understood to refer to themselves on the world-tube version of transtemporal identity in order to unmask what I shall argue is an illusion ignorance. Firstly, let's look at what Saunders and Wallace have to say in support of the idea. Here is Wallace :

> According to Lewis's proposal, if at some stage in my future I am to undergo branching into two copies, then (timelessly) there are two people, and my current (pre-branching) person stages are shared by both of them.
>
> On the additional assumption that the correct referent of utterances and of mental states is a person at a time (rather than a person-stage) it follows that I am genuinely ignorant of my post-branching future. For when I say 'who will I become' that statement should actually be ascribed to two versions of me (one of whom will, post splitting, become each version of me). Since (as a consequence of any physicalist approach to mind) any thoughts and beliefs I have at a time supervene on my person-stage at that time and since the two versions of me share all person-stages

prior to branching it follows that it is impossible for the two versions of me to resolve their ignorance.

What are they ignorant about ? Not of course any propositional knowledge, but something more indexical (2005a, sec. 3.4)

This is much too quick because we need to know more about how an utterance of 'I' refers to the utterer for the idea to be coherent, especially as we're in a novel situation where a single vocal event is to be understood to express the utterances of more than one person. It's not what Wallace dubs the 'additional assumption' that it is persons who utter rather than person-stages which is the problem. It's the lack of any hint of a mechanism whereby each of Wallace's dual utterers secures an indexical reference to themself by the use of 'I' prior to fission. Wallace covers this lacuna by alluding to the 'ascription' of each utterance to their respective utterer but, as I shall explain, this idea of ascription trades on a mechanism of indexical reference which is unproblematic in non-branching contexts but which is inapplicable to branching. Symptomatic of Wallace's neglect of the distinction between the identification of speakers and of the referents of their utterances is his use of the term 'referent' in line 4 of the above quote and the word 'ascribed' in line 6.

Saunders' and Wallace's recent paper on the idea reinforces the impression that they are failing to distinguish between the attribution of an utterance to an agent and the determination of the reference of the ostensibly self-referential term 'I' within that utterance :

We do better to attribute [*pace* Lewis] thoughts and utterances at t to continuants C at t. That is, thoughts or utterances are attributed to ordered pairs $\langle C,t\rangle$ or slices of persons $\langle C,S\rangle$, $S \in C$, not to temporal parts themselves. This is to apply whether of not there is branching. In the absence of branching we obtain the standard worm-theory view ; in the presence of branching, we conclude that there are two or more thoughts or utterances expressed at t, one for each of the continuants at that time.

Lewis ruled out this semantics peremptorily. Suppose continuants C_1 and C_2 share the temporal part S at t, and suppose C_1 dies shortly after the branching, whilst C_2 survives. Then, said Lewis, C_1 and C_2 'cannot share the straightforward commonsensical desire that he himself survive', because ……. [there follows a quote from Lewis (1976 :74)] …….

True enough if there is only one thought. But why not if there are two, if the referent of 'I myself', thought or uttered at time t (at temporal part S) is a continuant, as in the non-branching case ? (Saunders and Wallace 2007 :2)

I shall not detour into exegesis of Lewis here, but the fact is that Saunders' and Wallace's proposed semantics is incomplete. They provide no explanation whatsoever of how they pass from the idea of attributing utterances to continuants to the idea that 'the referent of « I myself » …. is a continuant'. To be clear about what is at issue here, generally in linguistics it is accepted that the attribution of an utterance to an agent, the person deemed to make that utterance, is distinct from whatever determines the referents of words and phrases uttered by that agent. What Saunders and Wallace are doing is effectively to maintain that this distinction need not be made in the case of a phrase like 'I myself'. They apparently assume that such a phrase self-evidently refers to the person to whom it is attributed. That may seem an innocuous assumption but in the context before us here it is crucial that it is not allowed to pass without some comment for Saunders' and Wallace's idea turns on this assumption. I shall explain in a moment that in non-branching situations where a world-tube analysis of transtemporal identity is adopted there is an unproblematic account of how the term 'I' refers to the person making an utterance of it and that that account is not available in branching contexts. What Saunders and Wallace need is the idea that utterances of 'I myself' made severally by overlapping continuant persons (and thus instantiated by single utterance tokens) severally refer to the persons uttering them. Saunders' and Wallace's proposed semantics needs to be supplemented by an account of how the reference of 'I myself' is determined for overlapping persons.

Providing such an account of continuant persons' self-reference in non-branching circumstances is straightforward. What is involved may be clearer if we look first of all at how ordinary indexical reference by a person to some object in their environment is understood to operate according to the world-tube view of transtemporal identity when fission is not involved. On the 'endurance' view of transtemporal identity a persisting object is 'wholly present' at all times in its history. The world-tube view denies this, taking a persisting object to be its history, an aggregate of temporal parts. On the endurance view a subject is wholly present at any time at which s/he makes an indexical reference, as is the environmental object to which s/he refers. As both referer and referent are wholly present at the time the reference is made there seems to be no problem in principle about there being an indexical relation between them at that time. In contrast, on the world-tube view it is only temporal parts of the referer and the referent which are present at the time the reference is made.

But there is no real difficulty for the world-tube theorist here. An utterance is made at a time by a person and the utterance is tokened by an event, usually vocal, which is associated with a stage which is a temporal part of that person's body according to world-tube theory. Suppose that in a non-branching context René, faced with an apple, says 'That apple is green' at time T. At time T a stage of René's body is associated with a vocal event

which is understood to be a token of ‘That apple is green’. At time T there is an apple-stage which is appropriately related to the body-stage associated with the vocal event and that apple-stage is a temporal part of a single apple. That's how René succeeds in indexically referring to an apple: there is an appropriate juxtaposition of the utterance token associated with a stage of René's body and a stage of the indicated apple. The idea brings to mind the image of a chromosome pair, touching in the middle: the world-tube subject successfully refers to a world-tube object at a time because stages of each world-tube are in an appropriate relation to each other at that time.

Now go on to the non-branching case where René says ‘This is my body’, an unusual statement, but we would generally take it to be perfectly intelligible. He might stub a finger at his chest for gestural emphasis but that would be strictly unnecessary, René’s use of ‘this’ would be enough to indicate the body in question. That is because the site of the token of ‘this’ is a body-stage which is a stage of a unique world-tube body. Like René’s reference to the apple, his reference to his body picks out a unique world-tube referent. But what of René’s reference to himself ? Here again the reference has to go via his body, there is nothing else which can provide evidence of which person René is, as is illustrated by everyday expressions such as ‘I’m over here !’.

For René’s utterance of ‘This is my body’ to be true the body picked out by the his use of ‘this’ has to be the body belonging to the person who is making the whole utterance. Clearly, the body-stage which is the site of the whole utterance is the very same as the body-stage which is the site of the utterance of ‘this’ and, thanks to the non-branching context, the body-stage in its role as determining a referent of the use of ‘this’ determines the same world-tube object as the body-stage in its role as determining the person who is the source of the utterance.

However, things do not go so smoothly for Lewis’s world-tube view of personal identity in branching contexts. To spell out why, let us return to the case of the Hydras. According to the Saunders/Wallace proposal both HydraUP and HydraDOWN at time T, prior to fission, can truly say ‘I am either HydraUP or HydraDOWN but I don't know which’. Both the Hydras say this severally at the same time since a single utterance token tokens two utterances, one made by HydraUP and the other made by HydraDOWN. Saunders and Wallace require that the Hydras are able to use ‘I’ in the everyday way in which we understand it, so that HydraUP refers to HydraUP when she uses ‘I’ and HydraDOWN’s ‘I’ refers to HydraDOWN. That must imply that both the Hydras can successfully indexically refer to their own bodies, since, as we saw with René, bodies are all we have to go on in determining which utterance of ‘I’ refers to which person.

But HydraUP and HydraDOWN cannot each indexically refer to her own body via an utterance of ‘This is my body’ which has a single token sited in a single body-stage at time T, prior to branching. For that single body-stage is common to the world-tube bodies of both HydraUP and HydraDOWN. Why should the ‘this’ in

HydraUP's utterance of 'This is my body' be understood to refer to HydraUP's body rather than to HydraDOWN's ? There is no reason. And if neither of the Hydras can secure reference to their own bodies then neither can secure reference to herself via an utterance of 'I'.

What is emerging here is that an utterance of 'I' in Lewisian contexts of multiple utterance cannot straightforwardly be assumed to refer to the utterer. It would not be good enough to say that this points to exactly the ignorance for which Saunders and Wallace want to argue because the breakdown of indexical reference which I have described simply makes the Hydras' statements of 'I am either HydraUP or HydraDOWN but I don't know which' unintelligible. Such an utterance would be as unintelligible as an utterance of 'That is green' in a context lacking any basis for a mechanism whereby the use of 'that' involved an indexical reference to a specific object.

This shows up the problem with Saunders' and Wallace's talk of 'attribution'. To attribute utterances of 'I' which share a unique token each to their respective utterer is simply to attribute agents to those utterances. That provides no warrant at all for supposing that the referents of those utterances of 'I' are the agents who do the uttering. To go on to assign the agent as referent is simply to assert that utterances of 'I' self-refer even in contexts of Lewisian multiple utterance without further ado. But that would be a substantive assumption on Saunders' and Wallace's part for which they give no justification and which leaves out of account altogether any mechanism by which the required indexical reference is secured even though such a mechanism is straightforwardly available for the world-tube view of transtemporal identity in non-branching contexts.

Might Wallace wish to appeal to his linguistic argument for charitable interpretation here in order to justify the further assumption which I have revealed? The argument could be that if we in fact inhabit an Everettian multiverse then all utterances of 'I' would fail to refer to the utterer on a world-tube view of personal identity and so we had better, out of linguistic charity, allow that they do so refer. But this would be to neglect that the world-tube view is not the only interpretation of transtemporal identity which can cope with branching. Since 1996 there has become available Ted Sider's 'stage theory' which can embrace the idea of continuant identity through branching without involving the concept of multiple utterance (Sider, 1996, 2001). According to stage theory persons are stages, not aggregates of stages, and so any utterance at a time has a token which is associated with the unique body, itself a stage, which is the body of that person at that time. Thus even if we do inhabit an Everettian multiverse utterances of 'I' can be understood to indexically refer to the utterer by the straightforward indexical mechanism of a token of a single utterance being appropriately associated with the body of a single person.

For readers not familiar with it, here is Sider's idea, which was not itself motivated by concerns about branching. Sider adapted Lewis's concept of modal counterparts to introduce the idea of temporal counterparts.

According to Lewis I have modal counterparts who are persons with blond hair in various 'possible worlds'. For any one of those modal counterparts I am not that person but I bear the relation MIGHT HAVE BEEN to that person. According to Sider I have past temporal counterparts who scrumped apples. For any one of those past temporal counterparts I am not that person but I bear the relation WAS to that person. If I am about to make a spin measurement in the Everett multiverse in the manner of Hydra then, according to Sider, I have future counterparts who see UP and future counterparts who see DOWN. For any one of those future counterparts I am not that person but I bear the relation WILL BE to that person, so I will be a person seeing UP at time T+ and I will be a person seeing DOWN at time T+. Those distinct future counterparts of mine are distinct persons (Sider, 2001 : 201).

Saunders and Wallace might wish to argue that Siderian transtemporal identity is not suitable if we inhabit an Everettian multiverse, that we are forced to accept Lewisian identity and that therefore, out of linguistic charity, we should generally interpret utterances of 'I' to refer to the utterer even though those utterances would be multiple in Lewis's sense. That would be a substantive argument which would need to be brought into play to support Saunders' and Wallace's proposal that a modified Lewisian semantics can motivate the idea of uncertainty of outcome prior to branching. Furthermore, it would imply that there is a more intimate connection between the metaphysics of identity and the argument from linguistic charity than Wallace appears to recognise in his writings to date.

In sum, the idea that Lewis's analysis of personal fission can be used to ground a notion of ignorance-based uncertainty prior to Everett branching appears to be inadequate unless it can be backed up by more extensive arguments than have been given so far. So Saunders' and Wallace's claim to have solved the 'incoherence problem' of the Everett interpretation is premature. They state the incoherence problem as being the idea that the Everett interpretation 'can give no meaning to the notion of uncertainty' (2007 :1). Bear in mind that Saunders' and Wallace's (2007) is concerned with establishing pre-measurement uncertainty and that it may be that the Everett interpretation can be rendered intelligible by appeal to a concept of post-measurement uncertainty such as Lev Vaidman's (1998). Wallace himself acknowledges such a possibility (2005a, sec. 4.2). [2]

References

Lewis, David, 1976: Survival and Identity. In Amelie O. Rorty (ed.), *The Identities of Persons*, University of California Press. Reprinted in Lewis, *Philosophical Papers*, vol. 1, Oxford University Press, 1983.

[2] Peter Lewis has come to a similar conclusion by a different route (2006). My thanks to him for comments on a previous draft of this paper.

Lewis, Peter, J., 2006: Uncertainty and probability for branching selves. Forthcoming in *Studies in the History and Philosophy of Modern Physics*.

Parfit, Derek, 1984. *Reasons and Persons*. Oxford : Oxford University Press.

Saunders, S. & Wallace, D. 2007 : Branching and Uncertainty. Available online at http://philsci-archive.pitt.edu/archive/00003383

Sider, Theodore 1996. All the world's a stage. *Australasian Journal of Philosophy* 74: 433 -- 53.

Sider, Theodore 2001. *Four Dimensionalism*. Oxford: Oxford University Press.

Vaidman, Lev 1998. On Schizophrenic Experiences of the Neutron or Why We Should Believe in the Many-Worlds Interpretation of Quantum Theory. *International Studies in Philosophy of Science* 12 : 245-61. Available online at http://arxiv.org/abs/quant-ph/9609006

Wallace, David, 2005a: Epistemology quantized: circumstances in which we should come to believe in the Everett interpretation. Forthcoming in *British Journal for the Philosophy of Science*.. Available online at http://philsci-archive.pitt.edu/archive/00002368

Wallace, David, 2005b: Language use in a branching universe. Available online at http://philsci-archive.pitt.edu/archive/00002554