Toy Models for Retrocausality*

Huw Price†

February 22, 2008

1 Motivation

A number of writers have been attracted to the idea that some of the puzzling features of quantum mechanics might be manifestations of 'reverse' or 'retro' causality, at a level underlying that of the usual quantum description. The main motivation for this view stems from EPR/Bell phenomena, where it offers two virtues. First, as was noted by its earliest proponent,[1] it has the potential to provide a timelike decomposition of the nonlocal correlations revealed in EPR cases – i.e., as we would now put it,[2] for the violation of the Bell inequalities in the quantum world. Second, Bell's derivation of his famous inequality depends explicitly on the assumption that hidden states do *not* depend on future measurement settings – so that its violation simply *invites* a retrocausal explanation, at least from the point of view of anyone who has already been bitten by the retrocausal bug.

Most people working in the foundations of quantum mechanics remain resolutely unbitten, however. It is common for the retrocausal option to be ignored altogether, or, as in this rather careful recent survey article, relegated to the footnotes with other *unmemorabilia:*

> To be scrupulous, there are perhaps four other ways [i.e., other than nonlocality] that the correlations in [an EPR-Bohm] experiment could be explained away. (1) One could simply 'refuse to consider the correlations mysterious'. (2) One could deny that the experimenters have free will to choose the settings of their measurement devices at random, as required for a statistically

*This note is based on a talk given at workshops at the University of Sydney and at Griffith University, Brisbane, in November, 2007. The slides for the Griffith University version of the talk are available online here: http://www.usyd.edu.au/time/price/preprints/RetroTalkGriffithNov07.pdf. I am grateful to John Cusbert, Pete Evans, Eric Cavalcanti and other participants in those workshops for helpful feedback, and to Steve Weinstein and especially Ken Wharton, for much helpful discussion since then. I am also indebted to the Australian Research Council and the University of Sydney, for research support.

†Centre for Time, University of Sydney; email: huw@mail.usyd.edu.au.

[1]The view was first proposed by Olivier Costa de Beauregard (1911–2007), a student of Louis de Broglie, whose first publication on the subject is (Costa de Beauregard 1953). Prof. Costa de Beauregard (2005) reported that he had proposed the idea several years earlier, in 1947, but that de Broglie had forbidden him to publish it, until Feynman's work on the positron gave some respectability to the idea of "things going backwards in time". For more recent retrocausal proposals, see the references in (Sutherland 2006).

[2]The qualification is necessary because Costa de Beauregard's version of the proposal pre-dates Bell's work by more than a decade, of course.

> valid Bell-experiment. (3) *One could entertain the idea of backward-in-time causation*. (4) One could conclude that ordinary (Boolean) logic is not valid in our Universe. I do not consider these escape routes because they seem to undercut the core assumptions necessary to undertake scientific experiments. (Wiseman 2005, my emphasis)

What can a fan of quantum retrocausality do at this point, to try to bring the proposal out of the footnotes and onto the main page? Well, there are two obvious strategies. The first is to construct explicit theories and models of quantum phenomena, embodying retrocausal principles. Various proposals of this kind are in the literature.[3] The second is to explore the conceptual foundations of the proposal – e.g., to examine the basis of our ordinary causal intuitions, in order, perhaps, to uncover some deep-seated errors in reasoning, underlying intuitive objections to retrocausality. Again, some work of this kind is in the literature – see, e.g., (Price 1996).

This note introduces a third strategy, which offers a promising complement to the other two, in my view. This third strategy aims to investigate retrocausality in general – and hopefully, eventually, *quantum* retrocausality – by developing simple 'toy models', to explain and elucidate its characteristics, and to explore its potential and peculiarities.

1.1 Playing with models

Much of the inspiration for this project comes from (Spekkens 2004), who proposes and investigates a 'toy theory', as he calls it, to explore the issue as to which of the distinctive features of quantum mechanics might be explained by the hypothesis that the quantum description is 'epistemic', rather than 'ontic'. By analogy, one aspect of my third strategy – admittedly, one that I make almost no progress with in this note – is to use toy models to explore the question as to what quantum-like phenomena retrocausality might in principle explain. The other aspect is to use such models as intuition pumps, or teaching aids, for clarifying and motivating the unfamiliar ideas involved in retrocausal proposals – e.g., for looking for latitude in what Wiseman (*op. cit.*) termed 'the core assumptions necessary to undertake scientific experiments.'

The toy model described here shows how something that 'looks like' retrocausality can emerge from global constraints on a very simple system of 'interactions', when the system in question is given a natural interpretation in the light of familiar assumptions about experimental intervention and observation. It yields nothing of a distinctively quantum nature, except a crude form of nonlocality. I present it in the hope that it may turn out to be a stepping stone to something more interesting, and especially in the hope that it will help to explain *what the game is,* to people who still find retrocausality as unattractive a response to the quantum puzzles as fatalism, non-Boolean logic, or a shrug of the shoulders.

[3] See, e.g., (Sutherland 2006) and the references therein, and (Wharton 2007).

2 The Helsinki model[4]

The Helsinki model is defined by the following ingredients and principles:

- There are two kinds of primitive nodes, each the inverse of the other under reflection around the horizontal axis, and each comprising a meeting-point of three edges. If we interpret the edges as 'particle world-lines', then the nodes represent two kinds of primitive 'interaction': 'pair production' and 'pair annihilation'. (See Fig. 1)
- Each edge has one of three 'flavours', A, B or C.
- Each node must be strictly **inhomogeneous** – i.e., comprising three edges of *different* flavours – or strictly **homogeneous** (three edges of the *same* flavour).
- Pair production and pair annihilation must alternate, when the primitive nodes are linked together.
- *Successive* homogeneous nodes are prohibited. (See Figs. 2 & 3)

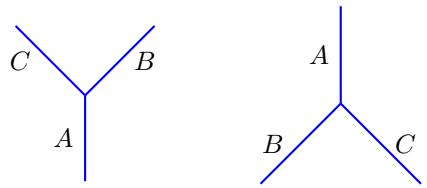

Figure 1: The two basic 'interactions'.

2.1 Adding 'time evolution', 'preparation' and 'observation'

The bare dynamics of this model is 'up-down' symmetric – or *time*-symmetric, if we treat up-down as a temporal axis (Fig. 4). Given such a temporal interpretation, however, then it is very natural to imagine we can *control* the inputs and *read off* the outputs, as in Fig. 5. Here yellow circles represent 'interventions', or 'preparations' (values we can 'choose to assign'); green squares represent 'observations' – values we simply 'read off'; and the wavy lines represent the 'hidden' sectors, that we can't directly control or observe. Note that the two pair annihilations in Fig. 5 provide 'measurements' of the hidden sectors, in the sense that if we know one input and the output, the rules uniquely determine the value of the second ('hidden') input.

[4] When I first presented this model at a workshop in Sydney in 2007, I explained that I had two reasons for calling it the Helsinki model: (i) I thought of retrocausality in QM as an elegant rival to Copenhagen (though somewhere in the same neighbourhood, in some respects); and (ii) the model itself first occurred to me while I was stuck between flights at Helsinki airport, a couple of months previously. Contrary to an interjection at that point by my student, John Cusbert, the name has nothing to do with "control by the Finnish state."

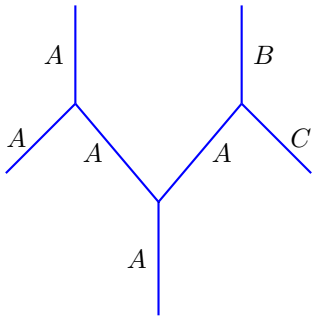

Figure 2: **Disallowed** – repeated homogeneous nodes.

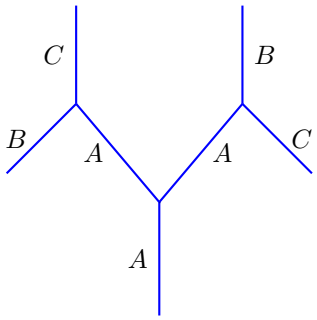

Figure 3: **Allowed** – no repeated homogeneous nodes.

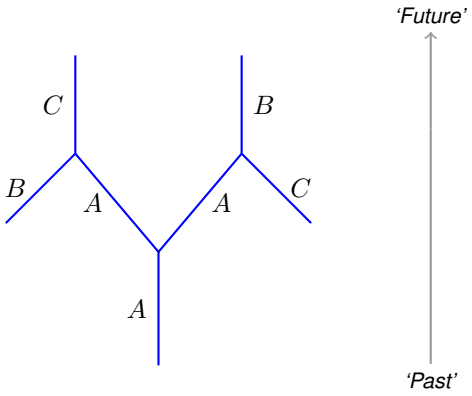

Figure 4: Adding a 'time axis'.

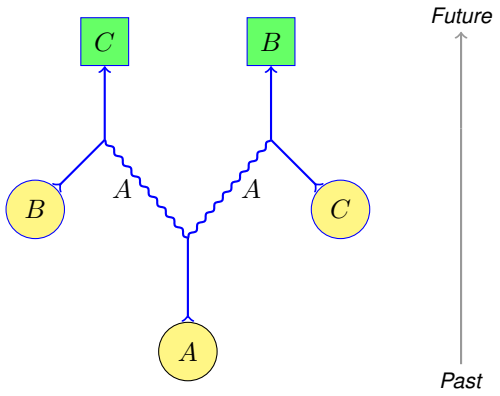

Figure 5: Adding 'preparations' and 'observations'.

For future reference, let's also emphasise that the direction of causation has been 'put in by hand', in this model, by our stipulation of what we can control. (It is certainly isn't given to us by the basic rules!) Our next tasks are (i) to explain what retrocausality amounts to, when the direction of causation is simply put in by hand in this way; and (ii) to show that the model requires retrocausality.

3 Reverse causation *v.* retrocausation

Since the direction of causation is put in by hand, we could put it in 'backwards', as in Fig. 6. Call this *reverse* causation: it corresponds to what causation looks like from the point of view of a creature whose time-sense is the reverse of our own. Since the Helsinki model is trivially time-symmetric (with the temporal interpretation we've given it),

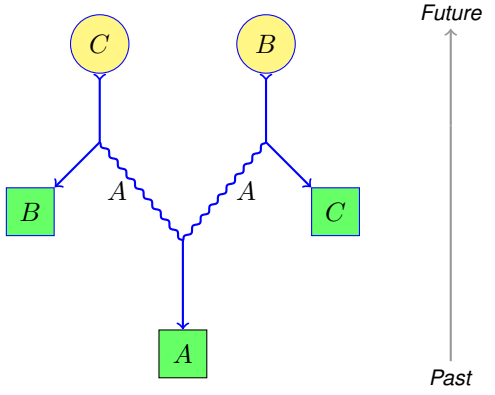

Figure 6: **Reverse** causation – interventions 'from the future', observations 'to the past'.

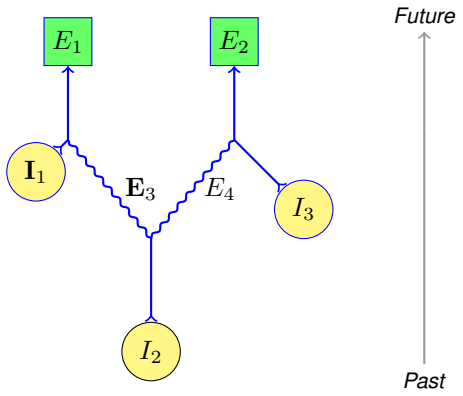

Figure 7: **Retro**causation – an intervention $\mathbf{I}_1$ may act 'backwards' on $\mathbf{E}_3$.

and the causal arrow is simply put in by hand, it is no big surprise to learn that it could be put in with the opposite orientation. And if this is what retrocausality amounted to, it could hardly be big news, surely, if we applied it to quantum mechanics?

Quite so, and the interesting case is entirely different. (One of the virtues of the Helsinki model is that it displays the difference so clearly.) The interesting case is when ordinary interventions ('from the past') make a difference *prior to the intervention* – e.g., in the notation of Fig. 7, if the choice of the 'measurement setting' $\mathbf{I}_1$ affects the 'hidden state' $\mathbf{E}_3$. (Here, think of the 'I_i' and 'E_j' as variables, representing the values of the three inputs, or *Interventions,* and some (potential) *Effects,* respectively. Each variable is restricted to the three values A, B or C, of course, by the rules of the model. The position of the input node labelled '$\mathbf{I}_1$' is intended to indicate that the choice of the value of $\mathbf{I}_1$ can be made *after* the time of the central 'pair production'.) This kind of influence – when the choice of $\mathbf{I}_1$ makes a difference to $\mathbf{E}_3$ – is what I want to call *retro*causation.

Unlike the case of reverse causation, which we can simply put in by hand – just a different choice of hand, so to speak – it is far from obvious that the Helsinki model involves retrocausation. To show that it does in fact do so, we need to investigate the patterns of correlations between inputs and hidden states allowed by the rules of the model. What we are looking for is a case in which a change in the left or right-hand input variables *requires* a change in the hidden state.

4 Retrocausality in Helsinki

To reveal the retrocausality in the Helsinki model, let's first consider the admissible three-input interactions (as in Fig. 5, for example). Exploiting the obvious symmetries of the model, there are effectively only four different choices of the three inputs. Writing the choice of inputs shown in Fig. 5 as '$B_A C$', for example, the four possibilities are $A_A A$, $A_A B$, $B_A B$, and $B_A C$ itself. For each of these choices of inputs, we want to know which of the nine possible hidden states – i.e., $\langle AA\rangle$, $\langle AB\rangle$, $\langle AC\rangle$, $\langle BA\rangle$, $\langle BB\rangle$, $\langle BC\rangle$,

$\langle CA\rangle$, $\langle CB\rangle$ and $\langle CC\rangle$ – are compatible with that choice. The fact that we have restricted ourselves to the case in which the central input is A immediately excludes most of these hidden states: the only admissible possibilities are $\langle AA\rangle$, $\langle BC\rangle$ and $\langle CB\rangle$. (The notation '$\langle XY\rangle$' is intended to indicate that X is the flavour of the hidden edge on the left, and Y that of the hidden edge on the right.)

This gives us only twelve cases to consider – four choices of inputs, and three hidden states for each – and the results are summarised in the State Table in Fig. 8. Note in particular that the inputs ${}_A A_A$ and ${}_A B_A$ *exclude* the hidden state $\langle AA\rangle$. This is the key to the model's retrocausality.

	$\langle AA\rangle$	$\langle BC\rangle$	$\langle CB\rangle$
${}_A A_A$	✗	✓	✓
${}_A B_A$	✗	✓	✓
${}_B A_B$	✓	✓	✓
${}_B A_C$	✓	✓	✓

Figure 8: The State Table.

4.1 Retrocausality revealed

Consider the case shown in Fig. 9. If either of the 'measurement settings' (i.e., the left or right inputs) were an A, as in Fig. 10, then the hidden state *couldn't be* $\langle AA\rangle$. (In the case shown in Fig. 10, with input A on the right, the two possibilities are a hidden state $\langle BC\rangle$ with left and right outputs both B; or a hidden state $\langle CB\rangle$ with left and right outputs A and C.) So in any actual case of the kind shown in Fig. 9, the hidden state depends *retrocausally* on the fact that neither 'observer' chose to input the measurement setting A rather than the measurement setting B. (As in Fig. 7, we could easily vary the position of the input nodes, to make it clear that the choice of measurement setting does not have to be made until *after* the pair production that produces the hidden state.)

Note also – comparing Fig. 9 and Fig. 10 – that the *output* on the left depends on the measurement setting on the right. If the actual case is as shown in Fig. 9, then again we have a counterfactual dependency, apparently: if we *had* chosen the input A on the right, we would have obtained either the output B or the output A on the left, rather than the output A. So we also have a kind of *nonlocality.*

5 Is the model consistent?

At this point there are some questions we might raise about the consistency of the model as a whole. For one thing, we might wonder whether are there larger systems constructed according to the same rules in which some choice of inputs allows *no* consistent assignment of outputs? The answer to this question seems to be 'no'. For suppose the contrary, and let N be the minimum length for the maximal string of sequential nodes in

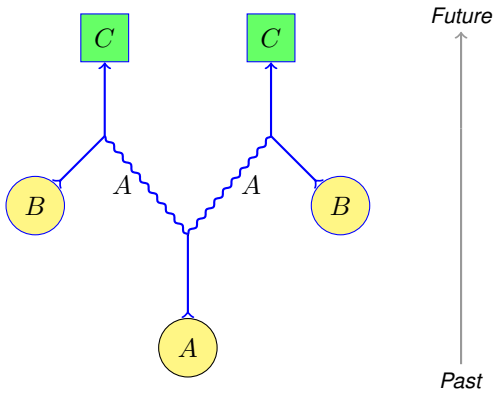

Figure 9: Hidden state $\langle AA \rangle$ is possible with these inputs.

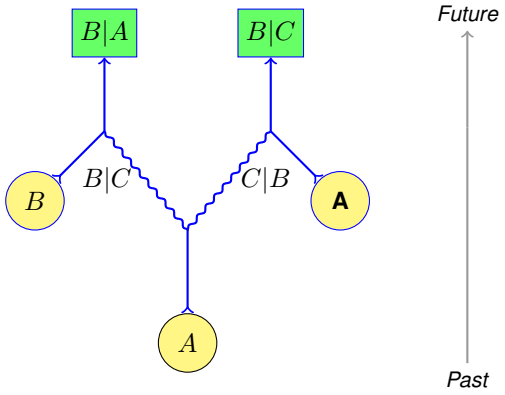

Figure 10: A different hidden state is enforced by the change of right input.

such an inconsistent structure. If $N > 1$, we could obtain a shorter inconsistent structure by choosing an inconsistent structure of length N, removing its lower-most level, and supplying by hand to the next level the inputs otherwise supplied by lower-most level. But this would contradict the assumption that N is the minimum length for such a structure, so $N = 1$. But there is no such system of length 1, apparently, and so no system of greater length, either.

This consistency property means that in interpreting the model in terms of our intuitive ideas of intervention, control and observation, we don't need to impose any restrictions on the ' free choices' of our toy physicists, in order to preserve consistency. This is an interesting result, especially in the light of the fact that two of the standard concerns about retrocausal models are that they might conflict with free will, and/or lead to inconsistencies or paradoxes of some kind.

5.1 Causal loops

We can take this concern with consistency under the natural interpretation a stage further, by allowing our toy physicists a freedom real physicists have in the case of EPR/Bell experiments, namely, to perform the two measurements at sufficiently different times, so that the result of one can be allowed to influence the setting of the other (by an ordinary 'classical' causal channel). This possibility is a recognised source of potential causal loops, in retrocausal models of EPR/Bell situations; see, e.g., (Berkovitz 2002).

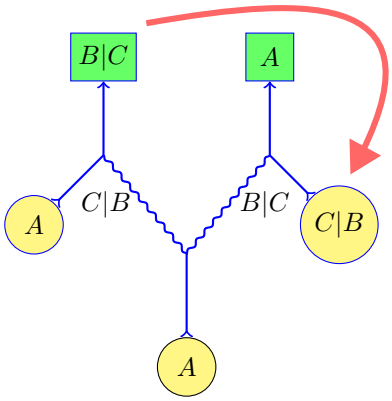

Figure 11: A causal loop? Left output controls right input.

Let's represent this possibility by adding to our diagrams the kind of causal link represented by the red arrow in Fig. 11. Keep in mind that despite the way it is depicted in Fig. 11, this is not to be thought of as a *retrocausal* influence. (Imagine the diagram elongated on the right, so that the right input actually occurs *after* the left output.) Keep in mind also that in this version, the new causal link lies in the classical realm of our toy physicists – it isn't part of the model itself. (As the model stands, the main obstacle to incorporating it within the model is the requirement that the two kinds of node must alternate – otherwise, we could simply make the output of the pair annihilation on the left an input of the pair annihilation on the right, eliminating the 'external' red arrow altogether.)

In the case shown in Fig. 11, the output B on the left produces input C on the right, and the output C on the left produces input B on the right. There is a consistent assignment of hidden states and right output in either case, showing that the constraint admits two consistent solutions (with the given choice of left and centre inputs – i.e., A in both positions).

Generalising this case, consider the three possible ways in which a left output B can fix a right input, as in Fig. 12. Again, all three cases allow a consistent assignment of the right output. Exploiting the symmetries of the model once more, this is sufficient to demonstrate that *whenever* the left and centre inputs are the same, *any* set of left-output-to-right-input constraints allows at least one consistent assignment of hidden states and right output – i.e., no such constraint can 'shut the system down'.

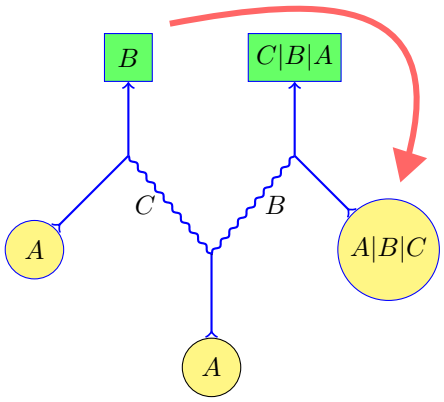

Figure 12: Generalising the previous case.

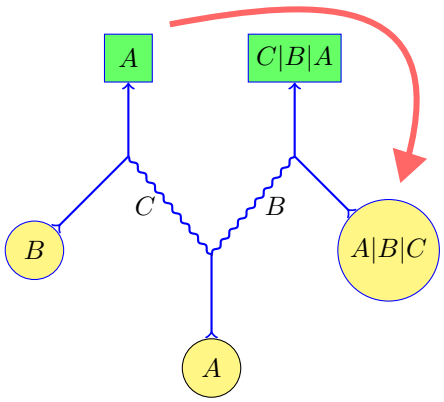

Figure 13: A new case, with different inputs in left and centre positions.

This leaves the cases in which the initial and left inputs are different. Here, consider, for example, the three possible ways in which a left output *A* can fix a right input, as in Fig. 13. Again, all three possibilities allow a consistent right output. And again, *any* set of left-output-to-right-input constraints allows at least one consistent assignment of hidden states and right output – again, no such constraint can shut the system down.

This kind of constraint is non-trivial, however. Fig. 14 shows a case in which its effect is to exclude a hidden state – $\langle AA \rangle$ – that would otherwise be permitted. So the Helsinki model is rich enough to show how this kind of causal loop can impose new constraints, without leading to inconsistency.

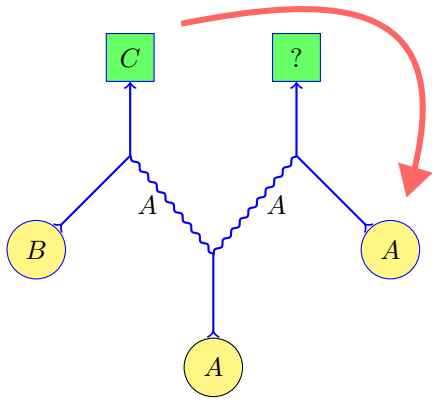

Figure 14: A substantial constraint.

6 Improving the model

I conclude with a wish-list of enhancements – further steps it would be interesting to be able to take in future iterations of the Helsinki model, or in something like it:

1. Adding probabilities, and showing that in virtue of the retrocausality, they are bound to have some of the characteristics of the probabilities associated with QM amplitudes – e.g., that probabilities of results of measurements cannot generally be regarded as probabilities of pre-existing states, regarded as *independent of the choice of future measurements.*

2. Developing an analogy between the standard QM state function and what we know in the Helsinki model *if we don't know the future measurement settings* – in other words, an analogy with the kind of epistemic 'coarse graining' of the Helsinki model which would be necessary to represent the state of knowledge of a physicist (or toy physicist) who wants to make predictions with respect to a range of possible 'next measurements'.

3. Hence connecting the Helsinki model, or some variant of it, to Spekkens' 'epistemic' toy models.

4. Investigating the nonlocality of Helsinki-like models, in search of the retrocausal toy modeller's Holy Grail: a model with Bell-like correlations without signalling.

I don't know to what extent such enhancements are possible, but I'll be pleased if the model inspires anyone to try to find out.

Abstract

A number of writers have been attracted to the idea that some of the peculiarities of quantum theory might be manifestations of 'backward' or 'retro' causality, underlying the quantum description. This idea has been explored in the literature in two main ways: firstly in a variety of explicit models of quantum systems, and secondly at a conceptual level. This note introduces a third approach, intended to complement the other two. It describes a simple toy model, which, under a natural interpretation, shows how retrocausality can emerge from simple global constraints. The model is also useful in permitting a clear distinction between the kind of retrocausality likely to be of interest in QM, and a different kind of reverse causality, with which it is liable to be confused. The model is proposed in the hope that future elaborations might throw light on the potential of retrocausality to account for quantum phenomena.

Bibliography

Berkovitz, J. 2002: On Causal Loops in the Quantum Realm. In T. Placek and J. Butterfield (eds.), *Non-locality and Modality,* Kluwer, 235–257.

Costa de Beauregard, O. 1953: Méchanique Quantique. *Comptes Rendus Académie des Sciences* 236, 1632.

Costa de Beauregard, O. 2005: Personal communication to Guido Bacciagaluppi and Huw Price, Bourron-Marlotte, France, 28.03.05.

Price, H. 1996: *Time's Arrow and Archimedes' Point.* New York: Oxford University Press.

Spekkens, R. W. 2004: In Defense of the Epistemic View of Quantum States: a Toy Theory. http://arXiv.org/abs/quant-ph/0401052

Sutherland, R. 2006: Causally Symmetric Bohm Model. Forthcoming in *Studies in the History and Philosophy of Modern Physics,* **39**, 2008. Preprint available online here: http://arxiv.org/abs/quant-ph/0601095v2

Wharton, K. 2007: A Novel Interpretation of the Klein-Gordon Equation. Preprint available here: http://arxiv.org/abs/0706.4075

Wiseman, H. 2005: From Einstein's Theorem to Bell's Theorem: A History of Quantum Nonlocality. http://arxiv.org/abs/quant-ph/0509061v3