Retrocausal Models for EPR

Richard Corry

2015

Abstract

This paper takes up Huw Price's challenge to develop a retrocausal toy model of the Bell-EPR experiment. I develop three such models which show that a consistent, local, hidden-variables interpretation of the EPR experiment is indeed possible, and which give a feel for the kind of retrocausation involved. The first of the models also makes clear a problematic feature of retrocausation: it seems that we cannot interpret the hidden elements of reality in a retrocausal model as possessing determinate dispositions to affect the outcome of experiments. This is a feature which Price has embraced, but Gordon Belot has argued that this feature renders retrocausal interpretations "unsuitable for formal development", and the lack of such determinate dispositions threatens to undermine the motivation for hidden-variables interpretations in the first place. But Price and Belot are both too quick in their assessment. I show that determinate dispositions are indeed consistent with retrocausation. What is more, I show that the ontological economy allowed by retrocausation holds out the promise of a classical understanding of spin and polarization.

NOTICE: this is the Accepted Maunsscript version of a work that was accepted for publication in *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, Vol 49, 2015. DOI: 10.1016/j.shpsb.2014.11.001

1 Introduction

One of the most troubling features of the Copenhagen interpretation of quantum mechanics is its assertion that some measurable properties do not have determinate values before they are measured. This indeterminateness is not only counterintuitive, it is also what makes the measurement problem so difficult for the Copenhagen Interpretation. To account for the fact that measurements always seem to have a determinate outcome, the Copenhagen interpretation

posits a special role for measurement in the dynamics of a system—to "collapse" the wavefunction such that the measured property takes on a determinate value. The problem is that there is no clear definition of what counts as a measurement, nor is there any explanation of why measurement should play this role. Hidden-variables interpretations seek to avoid these problems by insisting that the indeterminateness of quantum mechanics is merely epistemic. These interpretations claim that the formalism of quantum mechanics is incomplete in the sense that there are determinate "elements of physical reality" (to use Einstein, Podolsky, and Rosen's (1935) term) that have no counterpart in the formalism. The thought is that many of the puzzling aspects of quantum mechanics might arise from our ignorance of these "hidden" elements of reality. In particular, if the outcomes of all measurements are determined by such elements of reality, then we can interpret the collapse of the wavefunction as an epistemic issue; it represents an updating of our incomplete information about the world rather than a real change in the world from an indeterminate to a determinate state. In this way, hidden-variables interpretations hope to avoid ascribing indeterminateness to the world, and thereby hope to dissolve the measurement problem.

However, hidden-variables interpretations face a major problem: there are a number of No Hidden Variables theorems which seem to show that the assumption of hidden variables is incompatible with quantum mechanics. Of particular interest is John Bell's (1964) variation of the "EPR" thought experiment first presented in 1935 by Albert Einstein, Boris Podolsky, and Nathan Rosen (ironically, Einstein, Podolsky, and Rosen presented the original version as an argument that quantum mechanics must be incomplete). Bell showed that in this thought-experiment, hidden-variables interpretations make predictions that are at odds with accepted quantum mechanics. Later experimental results seem to uphold the predictions of quantum mechanics rather than hidden-variables (Aspect et al., 1982).

Physicists were quick to note that hidden-variables interpretations could be saved if they allow non-local interactions; in particular, if the interpretations include instantaneous action at a distance then they would not be inconsistent with quantum mechanics. Thus the conclusion was drawn that hidden-variables theories must be non-local. Physicists have long been suspicious of action-at-a-distance (despite such action playing a central role in Newton's theory of gravitation), and *instantaneous* action-at-a-distance is particularly problematic since it assumes an objective notion of simultaneity, in conflict with Einstein's theory of relativity. For this reason the EPR thought experiment is regarded by many as an important nail in the coffin of hidden-variables interpretations of quantum mechanics. This conclusion is a little unfair, however, since the EPR experiment also shows that non-hidden-variables interpretations like the Copenhagen interpretation must likewise involve a kind of instantaneous action-at-a-distance (this was essentially the point of Einstein, Podolsky, and Rosen's paper) and are thus no better off than hidden-variables theories in this respect.

There is, however, a way to make a local hidden-variables interpretation compatible with quantum mechanics. Bell's argument assumes that the values of the hidden variables are independent of the future settings of the experimental

apparatus. Indeed, this same assumption is made in all No Hidden Variables theorems. Thus, a number of writers have suggested that we can resolve many of the puzzles of quantum mechanics if we allow the possibility of *retrocausation*, whereby the properties of a system can be influenced by future events (see, e.g. Costa de Beauregard (1976), John Cramer (1980), Rod Sutherland (1983), David Miller (1996), Huw Price (1997), Phil Dowe (1997), and Ken Wharton (2007)).

Of course, retrocausation is itself rather counterintuitive and is often dismissed as involving paradox or problems for free will (see, for example, Bell's comments in Davies & Brown 1986, pp. 49-50). In response to such attitudes, Price introduced the strategy of investigating retrocausation by constructing "toy models" that can be used to explore and elucidate the possibilities of retrocausation. The first of these toy models—the *Helsinki model*—is designed to represent some very general features of retrocausation, and he expresses his hope that further models will be developed which capture more specifically quantum phenomena. In particular, he comments that a model that includes Bell-like correlations is the "retrocausal toy modeller's Holy Grail" (2008, p. 761).

This paper takes up Price's challenge and develops a retrocausal toy model of the Bell-EPR experiment. The model shows that a consistent, local, hidden-variables interpretation of the EPR experiment is indeed possible, and gives a feel for the kind of retrocausation involved. However, the model also makes clear a problematic feature of retrocausation: it seems that we cannot interpret the hidden elements of reality in a retrocausal model as possessing determinate dispositions to affect the outcome of experiments. This is a feature which Price (1997, p. 250) has embraced, however Gordon Belot has argued that this feature renders retrocausal interpretations "unsuitable for formal development" (Belot, 1998, p. 479), and the lack of such determinate dispositions threatens to undermine the motivation for hidden-variables interpretations in the first place. But Price and Belot are both too quick in their assessment. I will show that the retrocausal model is consistent with determinate dispositions so long as one accepts a particular view of the metaphysics of dispositions. I will also consider two variations of the original retrocausal model which allow for determinate dispositions even without this metaphysical assumption.

2 Modeling EPR

In the original EPR thought experiment, two particles are created in an entangled state, the two particles are then separated, and a measurement is performed on each of the separated particles. In what follows I will focus on Bohm's (1951) variation of the thought experiment. In this variation, we can set each of our measuring devices to make one of three different measurements, and each measurement gives one of two possible results (for example we might set the devices to measure the spin along three different axes). Call these three settings A, B, and C. The original EPR thought experiment is recovered if we allow only two settings and consider only situations in which the same setting is chosen for the

measurement of each particle. For convenience I will refer to Bohm-type EPR experiments simply as EPR experiments from now on.

The interesting features of the thought experiment derive from the facts that (i) the two particles are in an entangled state, meaning that they are not independent in some sense (to be discussed below); and (ii) measurements A, B, and C measure properties that are not simultaneously given determinate values in any quantum state description.

Following Price (2008), the models presented below focus on the causal structure of the thought experiment, and the relevant facts about entanglement and measurement are represented by placing constraints on the possible interactions. We begin, then, by noting that the EPR experiment involves three interactions: an interaction that produces a pair of entangled systems, and two measurement interactions. This structure is depicted in figure 1.

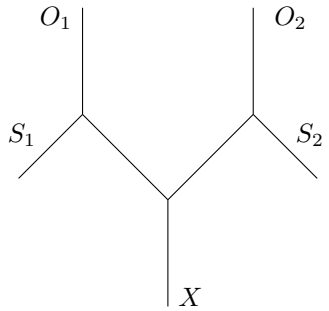


Figure 1: Causal model of the EPR experiment

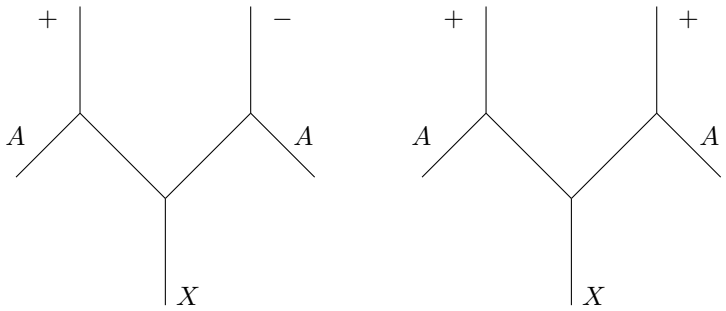
Here X is the device that produces the entangled state, S_1 and S_2 are the settings of the two measuring devices while O_1 and O_2 are the observed outcomes of the two experiments. We will let S_1 and S_2 each take values from the set $\{A, B, C\}$, representing the three measurement settings, while O_1 and O_2 each take values from the set $\{+, -\}$, to represent the two possible measurement outcomes. We will leave the possible values of X unspecified.

Vertical separation between nodes represents temporal separation, and we will stipulate that the future is towards the top of the page. Horizontal separation between nodes represents spatial separation. The unlabeled internal paths represent two physical systems that interact at one point in time, then move away from each other. Each of these systems is then involved in a measurement interaction at some later point in time. For convenience sake, I will refer to the leftmost of these systems as "particle 1" and the rightmost as "particle 2". In general, however, these systems need not be thought of as particles. We will ensure that our models are consistent with special relativity by insisting that all paths be null or timelike (and hence cannot represent systems traveling faster than the speed of light). Finally, let us stipulate that the two measurement interactions are simultaneous in the laboratory rest frame. Note that this last stipulation together with the restriction that paths be null or timelike imply that there can be no path directly connecting the two measurement interactions.

If we assume that at X we have a device for producing two particles in an appropriately entangled state, then quantum mechanics predicts—and experiment seems to confirm (Aspect et al., 1982)—the following two facts:

Fact 1 Whenever S_1 and S_2 have the same setting (regardless of what it is), O_1 and O_2 have opposite outcomes. So graphs like that in figure 2a are observed while graphs like that in figure 2b are not.

Fact 2 When S_1 and S_2 have different values, the observations O_1 and O_2 have different values close to 25% of the time.



(a) Observed (b) Never Observed

Figure 2: Outcomes of measurement

Our task is to construct a local hidden-variables model that is compatible with these two facts. At first glance, the task seems easy enough, but, as is now well known, it turns out not to be so straightforward; indeed the simple approach fails miserably.

3 A Simple Hidden-Variables Model

The most obvious explanation for the correlation described in Fact 1 is that the measuring devices are measuring determinate properties of the two entangled particles, and that these properties are perfectly anticorrelated during the entanglement interaction. This is the interpretation that Einstein, Podolsky, and Rosen (1935) argued for in their original version of the thought experiment. Now, according to this interpretation the two entangled systems have determinate properties for each of the three settings of the measuring devices. We can therefore represent each possible state of a particle as a triplet such as (+,+,−) which we read as representing the fact that the particle will produce an outcome of + if an A or B measurement is performed, and an outcome of − if a C measurement is performed. The general setup can then be represented as in figure 3, where a_1, b_1, c_1, a_2, b_2, and c_2 can each take the value + or −.

To produce the results described in fact 1, we simply add the following two constraints to the model:

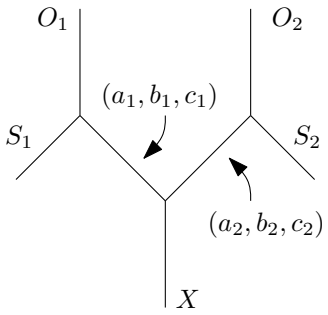


Figure 3: The hidden-variables model

Measurement Constraint

Measurement interactions, like that shown in figure 4a (or its left-right mirror image) must satisfy the following conditions:

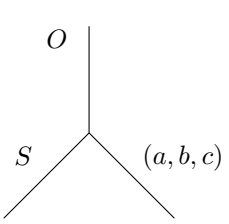
1. If $S = A$, then $O = a$
2. If $S = B$, then $O = b$
3. If $S = C$, then $O = c$

Entanglement Constraint

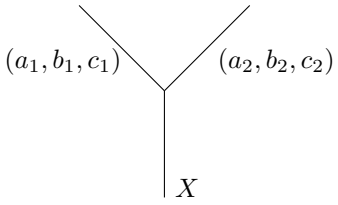
Entanglement interactions, as represented in figure 4b, must satisfy the following conditions:

1. $a_1 = -a_2$
2. $b_1 = -b_2$
3. $c_1 = -c_2$.

That is, we constrain the entanglement interaction to produce particle pairs whose properties are anticorrelated.



(a) A measurement interaction



(b) An entanglement interaction

Figure 4: Interactions in the model

The measurement constraint simply ensures that the measurement interactions really do result in a measurement of the values of the particle's properties. If the constraint were not satisfied, we would have no good reason to call this a measurement. Given the measurement constraint, the entanglement constraint is enough to ensure that any graph in which $S_1 = S_2$ will be such that $O_1 = -O_2$, in agreement with Fact 1.

Before discussing the problems with this model, it will be instructive to note a few of its features:

3.1 Probability

Although the measurement results always turn out to be anticorrelated, quantum mechanics predicts, and experiment confirms, that the actual result of any given measurement could be either "+" or "−", and the two possibilities will occur with some particular probability distribution. This fact can be accommodated in the model in two ways. (i) We could insist that each interaction is deterministic, but allow the input at X to have at least two possible values whose probabilities are appropriately specified. (ii) We could allow the entanglement interaction to be non-deterministic. In this case we would say that even given the same input at X, two entanglement interactions may produce different anticorrelated particle pairs. For example one interaction may produce the particles $(+,+,+)$ and $(-,-,-)$, while the second interaction produces the particles $(+,-,+)$ and $(-,+,-)$. Again the probabilities of each result would need to match the probability of the observed outcomes. There is no corresponding third option that the measurement interactions are probabilistic, since the measurement constraint implies that these interactions are deterministic.

For my purposes it will not be important which of these methods the model uses to generate probabilities. Thus, I will simply leave the values of X undefined and assume that the values on the internal paths satisfy an appropriate probability distribution.

3.2 Locality

The model represents each interaction as occurring at a spacetime point and involving only paths that intersect at that point. In particular, the constraints on a possible interaction involve only information that is present at that interaction node (and hence, information that is present at the relevant point of spacetime). Thus the simple hidden-variables model is local in the sense that there is no action at a distance. Furthermore, since we have insisted that paths cannot represent systems traveling faster than light, the model is local in the stronger sense that there is no faster than light transfer of information.

3.3 Free Will

The constraints on possible interactions described above are consistent with all combinations of settings for S_1 and S_2. Furthermore, any probability distri-

bution over the possible states produced at X will also be consistent with all combinations of settings for S_1 and S_2. Thus there is nothing in the model that restricts an agent's choice of settings for the measuring devices and as such the model does not rob experimental agents of free will.

3.4 Indeterminateness and the Measurement Problem.

Unlike the Copenhagen interpretation, the Simple Hidden-Variables Model assigns a determinate value to every measurable property of each particle. What is more, the measurement constraint ensures that the outcome of any measurement is determined by these values. Thus, there is no indeterminateness in the model that needs to collapse upon measurement, and so the Simple Hidden-Variables Model avoids the measurement problem. So, the Simple Hidden-Variables Model demonstrates the logical possibility of a local hidden-variables interpretation that is consistent with Fact 1, and which defuses the measurement problem.

4 The Bell Inequality

The problem with the simple hidden-variables model described above is that it conflicts with Fact 2. That is, the model conflicts with the fact that when S_1 and S_2 have different values, the observations O_1 and O_2 only have different results around 25% of the time.

To see the conflict, consider Table 1. All the possible experimental settings in which S_1 is not equal to S_2 are listed in the first column (Where AB represents $S_1 = A$ and $S_2 = B$, and so on) and the top row lists all combinations for the hidden variables that are allowed by the entanglement constraint. The table then indicates whether the observations at O_1 and O_2 are the same (S) or different (D).

	$(+,+,+)$ $(-,-,-)$	$(+,+,-)$ $(-,-,+)$	$(+,-,+)$ $(-,+,-)$	$(-,+,+)$ $(+,-,-)$	$(-,-,+)$ $(+,+,-)$	$(-,+,-)$ $(+,-,+)$	$(+,-,-)$ $(-,+,+)$	$(-,-,-)$ $(+,+,+)$
AB	D	D	S	S	D	S	S	D
BA	D	D	S	S	D	S	S	D
AC	D	S	D	S	S	D	S	D
CA	D	S	D	S	S	D	S	D
BC	D	S	S	D	S	S	D	D
CB	D	S	S	D	S	S	D	D

Table 1: Possible Measurement Outcomes

We see from the table that if the particles are in any of the six inhomogeneous states, we should expect to observe different results at O_1 and O_2 about 33% of the time. If the particles are in either of the two homogeneous states we

should expect to observe different results in all cases. Thus, no matter what distribution of states is produced in the experiment, the hidden variables model predicts that when $S_1 \neq S_2$ we should get different results at O_1 and O_2 at least 33% of the time. This fact about the simple hidden-variables interpretation was first pointed out by John Bell (1964) and is known as Bell's inequality. The Simple Hidden-Variables Model satisfies Bell's inequality, but Fact 2 states that Bell's inequality is violated by the real world.

5 A Retrocausal Hidden-Variables Model

Note that the constraints imposed in the simple hidden-variables model are stronger than strictly necessary to account for Fact 1. In order to ensure that the observed outcomes will be anticorrelated whenever the two measurement settings are the same, it is not necessary to ensure that all three properties of the two particles are anticorrelated. It is only necessary to ensure that the property actually being measured is anticorrelated. So, for example, when both measurement devices are set to A, we need only insist that the value of a_1 is anticorrelated with a_2; there is no need to also ensure that b_1 is anticorrelated with b_2, and c_1 anticorrelated with c_2. In order to incorporate this weaker constraint into the model, the entanglement interaction will need to "know" which measurement settings have been chosen so that it can ensure the correct variables are anticorrelated. The most straightforward way to give the entanglement interaction access to this information, whilst keeping the model local, is to add a variable representing the measurement setting to each of the internal paths. As we shall see, the result is a retrocausal model of the EPR interaction. This new model is depicted in Figure 5.

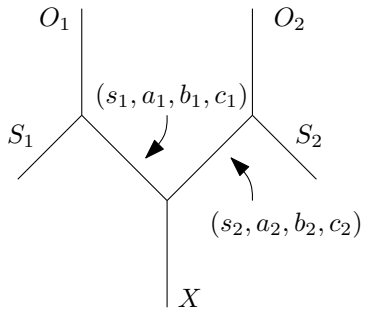


Figure 5: A retrocausal model

The weaker constraint can now be implemented by modifying the entanglement and measurement constraints as follows:

Retrocausal Measurement Constraint

Measurement interactions must satisfy the following conditions:

1. If $S = A$, then $O = a$
2. If $S = B$, then $O = b$
3. If $S = C$, then $O = c$
4. $s = S$

Retrocausal Entanglement Constraint

Entanglement interactions must satisfy the following conditions:

1. If $s_1 = s_2 = A$ then $a_1 = -a_2$
2. If $s_1 = s_2 = B$ then $b_1 = -b_2$
3. If $s_1 = s_2 = C$ then $c_1 = -c_2$.

Note that the Retrocausal Entanglement Constraint places no restriction on the value of the hidden variables when the measurement settings are different, and hence the model is consistent with any probability of anticorrelation of observation in such situations. In particular, then, the model is consistent with Fact 2 which asserts a 25% chance of anticorrelation when the measurement settings are different. Indeed, we can add a second entanglement constraint to ensure this outcome as follows:

Retrocausal Entanglement Constraint 2

Entanglement interactions must satisfy the following conditions:

1. If $s_1 = A$ and $s_2 = B$ then there is a .25 probability that $a_1 = -b_2$
2. If $s_1 = A$ and $s_2 = C$ then there is a .25 probability that $a_1 = -c_2$
3. If $s_1 = B$ and $s_2 = A$ then there is a .25 probability that $b_1 = -a_2$
4. If $s_1 = B$ and $s_2 = C$ then there is a .25 probability that $b_1 = -c_2$
5. If $s_1 = C$ and $s_2 = A$ then there is a .25 probability that $c_1 = -a_2$
6. If $s_1 = C$ and $s_2 = B$ then there is a .25 probability that $c_1 = -b_2$

The properties that are not being measured can have definite values, but the model need place no constraint on what these values are.

5.1 Retrocausality

Like the simple hidden-variables model the retrocausal hidden-variables model is local—both in the sense that it does not require any action at a distance, and in the relativistic sense that it does not require any faster than light transfer of

information. For, as in the case of the simple hidden-variables model, the constraints governing a possible interaction make reference only to the information available at that interaction node, and nodes are joined by paths that represent physical processes following null or timelike paths. Hence there is no violation of special relativity.

The second point to note is that it makes perfect sense to interpret the observations at O_1 and O_2 as measurements of pre-existing, determinate, properties of the system. Consider particle 1. As it travels along the internal path towards the measurement interaction, it has properties (s_1, a_1, b_1, c_1). The constraints we placed on the measurement interaction ensure that the observed outcome O_1 is always the same as the value of a_1 if the measurement setting is A, b_1 if the measurement setting is B, and c_1 if the measurement setting is C. Thus the interaction that produces outcome O_1 is a perfect measurement of the pre-existing value of the property it is set to measure. Similarly, O_2 is a perfect measurement of the pre-existing value of the relevant property of particle 2.

In this model, then, every measurement is the measurement of a perfectly determinate pre-existing property. In this respect, the model stands in contrast to the Copenhagen interpretation which states that measured properties in the EPR-experiment do not have determinate values until the measurement takes place (at which time the quantum state "collapses" to produce a determinate value).

What we have, then, is a local hidden-variables model that is compatible with the predictions of quantum mechanics in EPR type situations. That is to say, it is compatible with special relativity, it is compatible with a perfect anticorrelation of outcomes when identical measurements are made in an EPR situation, and it is compatible with the violation of Bell's inequality. Neither the simple hidden-variables theory, nor the Copenhagen interpretation can claim to satisfy all three of these desiderata. What is more, as a hidden-variables interpretation, the retrocausal model promises to relegate much of the weirdness of quantum mechanics to epistemology, rather than placing this weirdness out there in the world. So what is the cost of this interpretative paradise?

The most obvious cost of the model is that it seems to involve some kind of reverse causation. The value of $s_1$and the value of S_1 are perfectly correlated, and if this correlation is the result of a causal relation between $s_1$and S_1, then either the value of one determines the value of the other, or there is some common cause that determines them both. Our model does not include any such common cause, so in the model either $s_1$determines S_1or vice versa. But the only interaction between particle 1 and the left-hand measuring device takes place at a time after these variables have determinate values. And we can set up the experiment such that the entanglement interaction and the setting of the measurement apparatus are spacelike separated, so that there is no way to add a process connecting these events that does not travel faster than light. Hence at least one of the variables is determined by an interaction in its future. In particular, if we insist that the experimenter is free to choose any setting for each measurement device, then we must conclude that it is S_1that determines s_1.

Note that most causal interaction in the model can occur in the normal direction (so the experimenter's decisions can affect the later setting of the apparatus, measuring devices measure the state the particles had, not the state they will have). Indeed, if we apply the reasoning of the previous paragraph to the anticorrelation between a_1 and a_2 then we can conclude that there *must* be some determination in the normal direction, since particle 1 and particle 2 only interact before these variables have determinate values. Thus the model suggests a situation in which most causation happens in the normal direction, but in which there is some causation that goes the other way. Price (2008) calls this kind of feature *retrocausation*.

There are two ways that we might avoid this conclusion of retrocausation: First, we could modify the model to include an earlier common cause for $S_1$and s_1. Adding a common cause does not require any non-local interactions, so the new model would seem to have the benefits of the retrocausal model without the cost of retrocausation. However, this common cause would correspond to an unknown process that is able to consistently affect all possible mechanisms for setting the measurement apparatus (including affecting the apparent free choice of the experimenter). What is more, the common cause would have to act despite all attempts to shield the equipment. Peter Lewis (2006) dubs this the "hidden-mechanism conspiracy theory" of quantum mechanics, and argues that it is not only counterintuitive, but incoherent. I will not consider this possibility further.

The second way to avoid retrocausation in the model is to deny that the correlation between $S_1$and $s_1$is the result of a causal relation at all. It is a common suggestion that the correlations between measurement outcomes in the EPR experiment are not causal, the thought being that the measurements on the two particles are not really separate events, but constitute, rather, a "single, indivisible non-local event" (Skyrms 1984, p. 255, see also van Fraassen 1982, Fine 1989, and Hausman & Woodward 1999 for similar thoughts). However, it is far from clear that this reasoning could apply to the correlation between $S_1$and s_1. What is more, the whole point of the toy model is to provide a causal account of the correlation between measurement outcomes, so it would be perverse to now deny a causal explanation of the correlations within the model.

6 Does the Retrocausal Model Solve the Measurement Problem?

For the small cost of accepting a form of backwards causation, the retrocausal model provides us with a hidden-variables interpretation of EPR situations that is truly local. This locality is something that neither the Copenhagen interpretation nor the simple hidden-variables approach can claim, and hence the retrocausal model already seems to have one mark in its favor. But the motivation behind developing a hidden-variables model was that such models promise to

explain away many of the mysteries of quantum mechanics as artifacts of our ignorance. In particular, we might hope that—as a hidden-variables theory—the retrocausal model will describe determinate elements of reality that determine the outcomes of all possible measurements, and hence will allow us to avoid the measurement problem. Unfortunately, things are not as simple as we might hope.

Let us consider whether the model ascribes determinate values to all the measurable properties of the particles. In our model there are three measurements that can be performed on each particle, and each particle has three variables that are supposed to represent determinate values of the properties that these measurements measure. But simply having the right number of variables in the model is not enough. These variables must actually represent the appropriate properties, and this is where things get tricky. Let us call the property measured by an apparatus that is set to A, the A-property. Similarly, the B-property is what is measured by an apparatus set to B, and the C-property is measured by an apparatus set to C. Consider, now, the situation depicted in Figure 6, in which particle 1 is measured by a device that is set to A. The a_1-variable clearly represents the particle's A-property since the value of a_1 is the value that will be observed by the A-measurement. But what of the unobserved properties? In what sense do b_1 and c_1 represent determinate but unobserved B- and C-properties? The model places no constraint on the unobserved properties other than that their values are represented by "+" and "−", and so there is nothing in the model as it stands to ensure that the unobserved variables really represent A-, B-, or C-properties rather than some other, entirely irrelevant, properties (or no property at all). In order to ensure that all measur*able* (as opposed to measured) properties have determinate values, therefore, we need to add something to our interpretation of the model to ensure that the variables represent the same properties when unobserved as they do when they are observed (I will consider below the possibility that the unobserved properties need not have determinate values).

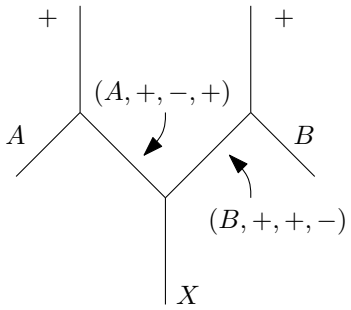


Figure 6: An allowed state

In the previous paragraph, we defined A-, B- and C-properties via their

causal roles in measurement.[1] Thus, we might hope to ensure that our variables always represent the appropriate A-, B- and C-properties by stipulating that the following conditionals are true whether the variables are observed or not:

Ca if a particle's a-variable has the value $+[-]$ then if a measurement of A were to be performed on that particle, the outcome would be $+[-]$.

Cb if a particle's b-variable has the value $+[-]$ then if a measurement of B were to be performed on that particle, the outcome would be $+[-]$.

Cc if a particle's c-variable has the value $+[-]$ then if a measurement of C were to be performed on that particle, the outcome would be $+[-]$.

The problem is that none of these conditionals can be true. To see why, consider again the situation depicted in Figure 6. We have seen that, in order to account for Fact 2, the a-variables cannot always be anticorrelated when one is not measured (and the same goes for the b- and c-variables), so situations like that shown in this figure must occur. But now let us now ask what would have happened if we had set the right hand device to A instead of B. Ca tells us that we would observe "+" on both sides, but this is inconsistent with Fact 1, which states that when the same measurement is performed on both particles, the results are always anticorrelated. Thus we cannot consistently interpret the variables as telling us what we would observe if we were to make observations of A, B or C.

What we have here is an example of a general issue with retrocausation that has been noticed (and embraced) by Price:

> Classically, it is natural to think of the state of a system as the sum of its dispositions to respond to the range of circumstances it might encounter in the future. But if the present state is allowed to depend on future circumstances, this conception seems inappropriate.(Price, 1997, p. 250)

Price concludes that in order to accommodate retrocausation this classical conception of measurements activating pre-existing dispositions "has to go", but he admits that it is "far from clear how to characterize what must replace it" (1997, p. 260). Gordon Belot is less sanguine about this issue, arguing that it renders Price's retrocausal interpretation "unsuitable for formal development" (Belot, 1998, p. 479). One might also worry that this problem defeats the purpose of a hidden variables interpretation. The main attraction of a hidden-variables interpretation, at least for authors like Einstein, is the hope that such interpretations allow us to see measurements as responding to determinate elements of reality and thereby avoid the measurement problem. But these elements of reality seem to have disappeared from the picture. As Belot quips, "Price's

[1] Shoemaker (1980; 2011), has argued that the causal profile of a property is essential to it. If he is right then our identification of A-, B-, and C-properties will succeed in any world. But even if the causal profile of a property is merely contingent, it is surely correct that, given the actual laws of nature, we can identify properties via their causal roles.

interpretation is the ultimate hidden variables interpretation. The only things missing are the variables." (1998, p. 479)

However, Price and Belot are too quick to conclude that pre-existing dispositions are incompatible with retrocausal models. For, as I will show, there is a way to interpret the variables of our retrocausal model as representing determinate dispositions.

6.1 Retrofinks

The use of conditionals like Ca, Cb, and Cc is a common strategy for characterizing dispositions. For example, we might understand the dispositional concept of fragility to imply that if a vase is fragile, then the following conditional is true:

Cf If the vase were struck, it would break.

However, it has long been known that such conditional analyses of dispositions are problematic. David Lewis (1997), for example, has us imagine a case in which a fragile vase is protected by a sorcerer. If the vase is ever struck, the sorcerer will immediately cast a spell to render it unbreakable. In this case, Lewis argues, the vase is fragile even though Cf is false. After all, says Lewis, before the sorcerer casts the spell the vase is intrinsically no different to any other fragile vase. The natural way to understand the situation, says Lewis, is that the vase is fragile before it is struck, but it loses this disposition as soon as it is actually struck. Lewis coined the term "finkish disposition" to describe a situation in which a disposition is lost as soon as the conditions are right for it to manifest itself.

Perhaps, then, we can understand the unobserved variables in the retrocausal model as representing something like finkish dispositions. That is, perhaps we can interpret "$a_2 = +$" as asserting that particle 2's A-property is disposed to produce the outcome $+$ when measured, but—like a finkish disposition—it will fail to have this disposition if such a measurement were actually performed. If, like many philosophers, we are happy to countenance finkish dispositions, then it seems there is nothing to stop us interpreting the unobserved variables as representing determinate properties of the particle.

There is, of course, a major difference between the standard finkish dispositions and those being conceived here. For a standard finkish disposition is one that is lost after it is triggered but before it has had time to manifest (e.g. after the vase is struck, but before it breaks). However, if we consider the situation in Figure 6 and ask what would have happened if particle 2's A-property were measured instead of its B-property, the answer is not that the particle would lose the disposition to produce outcome $+$ when it is measured, but rather, that the particle would not have had that disposition in the first place. The disposition is not finkish, but, we might say, *retrofinkish*.

So long as we are willing to accept the possibility of retrofinkish dispositions, we can interpret the retrocausal model as assigning determinant values to all measurable properties, and these values can be understood as dispositions

to produce determinant measurement outcomes. The result is a local hidden-variables interpretation that successfully avoids the measurement problem.

If you are willing to accept the possibility of both finkish dispositions and retrocausation, then it is not a big stretch to also accept retrofinks. Still, one might have worries about retrofinks. Some philosophers (e.g. Gundersen 2002 and Choi 2006; 2008) have denied the existence of even ordinary finkish dispositions, and their arguments would seem to rule out retrofinkish dispositions as well (indeed, their arguments would seem to show that particle 2 in Figure 6 has no value for its A-property). Furthermore, there does seem something worrying in the idea of a disposition that, as a matter of nomic necessity, cannot be manifest. For there seems to be some constitutive relation between laws of nature on the one hand, and dispositions on the other.[2]. It is interesting, therefore, to consider whether the retrocausal model can be varied in a way that avoids finkish dispositions. In what follows I will explore two such variations.

7 Retro-Copenhagen

One might think that it is not necessary to fuss around with retrofinks in order to get a solution to the measurement problem out of the retrocausal model. For surely we don't need to insist that there be determinate elements of reality determining every *possible* measurement outcome. Couldn't we avoid the measurement problem by merely insisting that the outcome of every *actual* measurement is determined by a determinate element of reality? Isn't it enough that no actual measurements ever collapse the wave-function, even if merely possible measurements could?

Since the measurement and entanglement constraints of the Retrocausal Hidden-Variables Model do not make reference to the unobserved properties of the particles, there is nothing to stop us interpreting these unobserved properties as being indeterminate, and we can do this without making any change to the constraints (and so still account for Facts 1 and 2). Let us call the interpretation that results the "Retro-Copenhagen" interpretation. Although this interpretation—like the Copenhagen interpretation—assigns indeterminateness to the world, the retro-Copenhagen interpretation assigns definite values to all the properties that are actually measured, and so might not involve any actual collapse. In fact, however, this interpretation of the retrocausal model suffers from its own version of the measurement problem.

To see this, consider what would happen if we performed two consecutive measurements on particle 1. In particular, suppose that we first perform a measurement with a device set to A, then perform a measurement with a second device set to B. The Retro-Copenhagen interpretation of our model tells us that before the A-measurement, particle 1 has a determinate A-property, but does

[2]The constitution relation might go either way. Lewis (1997) and Armstrong (1997, pp. 80-83), for example, have argued that dispositions are partly constituted by the laws of nature, while others, such as Ellis and Lierse (1994), and Bird (2007, pp. 43-64)have argued that the laws are constituted by dispositions)

not have determinate B- or C-properties. Our retrocausal model does not have the resources to tell us what happens to the particle after the A-measurement, but since we are about to perform a B-measurement, the natural way to extend the model would suggest that particle 1 should now have a determinate B-property which will be revealed by the B-measurement. But if the particle has a determinate B-property, then the natural extension of the Retro-Copenhagen interpretation would also suggest that the particle does not have determinate A- or C-properties. This situation is depicted in Figure 7.

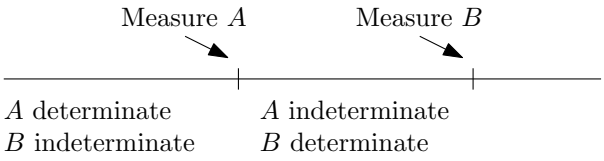


Figure 7: Consecutive measurements

There is a strong symmetry here between this interpretation of the retrocausal model and the Copenhagen interpretation. The Copenhagen interpretation tells us that properties may be indeterminate before they are measured, but guarantees that these properties have a determinate value after measurement. The Retro-Copenhagen interpretation, on the other hand, guarantees that measured properties have a determinate value before measurement, but implies that they may be indeterminate after measurement.

The symmetry between the Copenhagen interpretation and the Retro-Copenhagen interpretation carries over to the measurement problem as well. Before an A-measurement, the particle has determinate A-property, but indeterminate B- and C-properties, and similarly for other measurements. Why this correlation between the measurement to be performed and the determinate property? In order to preserve the free will of our experimenters when it comes to choosing the setting of the device, we must deny that this correlation is due to a common cause or that the choice of setting is determined by the determinate property. Thus we are left with the conclusion that it is the measurement interaction that determines which property has a determinate value, and this information then travels backwards in time (which is precisely why this is a retrocausal model). In essence, then, the Retro-Copenhagen interpretation tells us that when we make a measurement of A, we "collapse" the state to one in which the particle has a determinate A-property, and then propagate this state *backwards* in time. Furthermore, since this interpretation locates indeterminateness in the world, we are forced to interpret this collapse as making a real physical difference. The collapse determines which properties, out there in the world, have determinate values. Once again there is a symmetry between the Retro-Copenhagen and the Copenhagen interpretations. Both involve a physically significant collapse at the time of measurement, the difference being that the Copenhagen interpretation applies the collapsed state to times after the measurement, while the Retro-Copenhagen interpretation applies it to times before the measurement.

This interpretation of the retrocausal model, therefore, is no better at solving the measurement problem than is the Copenhagen interpretation.

8 A Simplified Retrocausal Model

The troubles we have encountered for the retrocausal model so far have all arisen from the unobserved variables in the model. If we insist that these variables take on determinate values, we are forced to accept nomologically-necessary retrofinks. If, on the other hand, we allow that these variables may be indeterminate, we are left with a model that is no better at solving the measurement problem than is the Copenhagen interpretation. But there is a third option: we can simply deny that there are any unobserved A-, B-, or C-properties.

To see how this third option might work, note first that the unobserved variables play no substantive role in our retrocausal model. In particular, they do not appear in any of the entanglement or measurement constraints. Thus it is possible to simplify the model by dropping the unobserved variables as follows.

First, rather than having four hidden variables for each particle we will have only two, s_i, which carries information about the experimental setting at S_i, and o_i which determines the outcome at O_i. s_i will take values from the set $\{A, B, C\}$ and o_i will take values from the set $\{+, -\}$. The complete model is shown in Figure 8:

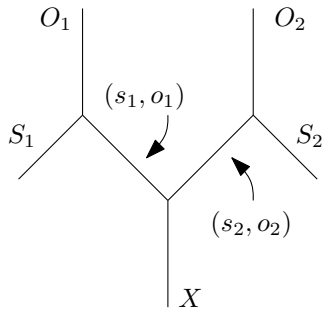


Figure 8: A simplified retrocausal model

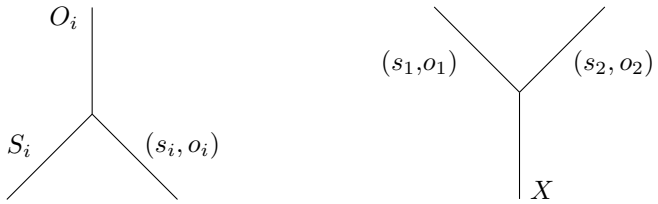
We now modify the measurement constraint and the entanglement constraint as follows:

> **Simplified Retrocausal Measurement Constraint**
> Measurement interactions like that shown in Figure 9a (or its left-right mirror image) must satisfy the following conditions:
>
> 1. $O_i = o_i$
> 2. $s_i = S_i$

Simplified Retrocausal Entanglement Constraint
Entanglement interactions, as represented in Figure 9b, must satisfy the following condition:

if $s_1 = s_2$ then $o_1 = -o_2$.

That is, we constrain the entanglement interaction to produce anti-correlated particles when the same measurement is being performed on each.



(a) Measurement (b) Entanglement

Figure 9: Simplified retrocausal interactions

These constraints are the weakest possible to ensure that when S_1 and S_2 are the same, the observations O_1 and O_2 will be anticorrelated, as required by Fact 1. In particular, these constraints impose no restriction at all on the possible outcomes at O_1 and O_2 in cases where S_1 differs from S_2. Thus we can recover the statistical result described in Fact 2, by positing that the entanglement interaction is 3 times as likely to involve states in which $s_1 \neq s_2$ and $o_1 = o_2$ as it is to involve states in which $s_1 \neq s_2$ and $o_1 \neq o_2$. Our simplified retrocausal model, therefore, is compatible with both Fact 1 and Fact 2.

But how do we interpret the simplified model? We could think of o_i as representing whichever of the A-, B-, or C- properties is being measured, so that the answer to the question "which property does o_i represent?" will depend on the values of the other variables in the model. But this is to take o_i as a mere notational convenience and accept that there are distinct A-, B-, and C- properties. If we are to avoid the problems associated with unobserved properties, we must deny that there are distinct A-, B-, and C- properties, and take our simplified model to give a complete description of all the relevant properties.

It turns out that there is a very natural and enlightening way to understand such a model. The EPR experiment involves the measurement of properties, like spin or polarization, which can only be measured with respect to a particular axis, and in each case these properties give rise to their own interpretive difficulties. To measure the spin of an electron, for example, we pass the electron between two strong magnets and find that the electron swerves either to the left or the right. This is exactly what we would expect to see if the electron were

spinning around an axis that is perpendicular to the line between the two magnets. If the electron were spinning clockwise around this axis, it would swerve left; if it were spinning anticlockwise it would swerve right. The problem with this understanding is that we can rotate the alignment of our magnets any way we like, and in each case the electron will behave as if it were spinning around an axis perpendicular to this alignment. In short, the electron behaves as if it were spinning around all axes simultaneously, which, of course, is impossible. Thus we are told that electron spin is a quantum property that is a bit—but only a bit—like the spinning of a top.

The simplified retrocausal model above, however, suggests a simple understanding of spin. Suppose that particle 1 is an electron and the A, B and C measurement are measurements of spin along three axes (each rotated 60° from the others). Then the variable $s_1$correlates with the axis of the measurement, while the the variable o_1 correlates with the spin of the electron with respect to that axis. So the electron is behaving as if it is spinning in direction o_1 around the axis s_1. What is more, the simplified model requires us to deny that the electron possesses any separate spin property related to axes that are not going to be measured. Thus the simplified model allows us to interpret the electron as literally spinning around a single axis. This straightforward classical understanding of spin is made possible by retrocausality. The axis around which the electron is spinning is determined retrocausally by the measurement interaction, while the direction of spin is determined by the entanglement interaction.

The Simplified Retrocausal Model brings out another feature of retrocausation: ontological economy. The existence of retrocausation would allow us, in some situations at least, to make do with fewer postulated properties than we would need without retrocausation.[3] In the case of our model, this ontological economy suggests an intuitive classical understanding of spin, and a similar understanding is available for polarization. This raises the intriguing possibility that retrocausation may be able to explain away the weirdness of these quantum properties more generally.

9 Conclusion

The Retrocausal Model and the Simplified Retrocausal Model each demonstrate the possibility of a local hidden-variables interpretation that is consistent with the facts of the Bell-EPR experiment, and which can successfully avoid the measurement problem. These models also show that retrocausation does not force us to abandon the idea that the state of a system is the sum of its dispositions to respond to the range of circumstances it might encounter. The full-blown Retrocausal Model requires us to countenance the metaphysics of retrofinks, but the Simplified Retrocausal Model does not require any controversial metaphysics beyond retrocausation. What is more, the Simplified Retrocausal Model suggests a classical interpretation of spin and polarization. If we are willing to

[3] Price (2012, Section 5.2) provides another example of the ontological economy of retrocausation.

accept retrocausation, then, the Simplified Retrocausal Model seems to hold out the promise of interpretive nirvana.

References

Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, *49*, 1804–1807.

Bell, J. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, *1*, 195–200.

Belot, G. (1998). Review of Time's Arrow and Archimedes' Point: New Directions for the Physics of Time. by Huw Price. *The Philosophical Review*, *107*(3), 477–480.
URL http://www.jstor.org/stable/2998455

Bird, A. (2007). *Nature's Metaphysics: Laws and Properties*. Oxford: Oxford University Press.

Bohm, D. (1951). *Quantum Theory*. London: Constable and Company.

Choi, S. (2006). The simple vs. reformed conditional analysis of dispositions. *Synthese*, *148*(2), 369–379.

Choi, S. (2008). Dispositional properties and counterfactual conditionals. *Mind*, *117*(468), 795–841.

Costa de Beauregard, O. (1976). Time symmetry and interpretation of quantum mechanics. *Foundations of Physics*, *6*(5), 539–559.
URL http://dx.doi.org/10.1007/BF00715107

Cramer, J. G. (1980). Generalized absorber theory and the einstein-podolsky-rosen paradox. *Phys. Rev. D*, *22*, 362–376.
URL http://link.aps.org/doi/10.1103/PhysRevD.22.362

Davies, P. C. W., & Brown, J. R. (1986). *The Ghost in the Atom: A Discussion of the Mysteries of Quantum Physics*. Cambridge: Cambridge University Press.

Dowe, P. (1997). A defense of backwards in time causation models in quantum mechanics. *Synthese*, *112*(2), 233–246.

Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, *47*(10), 777–780.

Ellis, B., & Lierse, C. (1994). Dispositional essentialism. *Australasian Journal of Philosophy*, *72*(1), 27–45.

Fine, A. (1989). Do correlations need to be explained? In *Philoophical Consequences of Quantum Theory*, (pp. 175–194). University of Notre Dame Press.

Gundersen, L. (2002). In defence of the conditional account of dispositions. *Synthese*, *130*(3), 389–411.

Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal markov condition. *The British journal for the philosophy of science*, *50*(4), 521–583.

Lewis, D. (1997). Finkish dispositions. *Philosophical Quarterly*, *47*(187), 143–158.

Lewis, P. J. (2006). Conspiracy theories of quantum mechanics. *The British Journal for the Philosophy of Science*, *57*(2), 359–381.
URL http://bjps.oxfordjournals.org/content/57/2/359

Miller, D. J. (1996). Realism and time symmetry in quantum mechanics. *Physics Letters A*, *222*(12), 31 – 36.
URL http://www.sciencedirect.com/science/article/pii/0375960196006202

Price, H. (1997). *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. Oxford: Oxford University Press.

Price, H. (2008). Toy models for retrocausality. *Studies in History and Philosophy of Modern Physics*, *39*, 752–761.
URL http://arxiv.org/abs/0802.3230

Price, H. (2012). Does time-symmetry imply retrocausality? how the quantum world says "maybe"? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *43*(2), 75–83.

Shoemaker, S. (1980). Causality and properties. In P. van Inwagen (Ed.) *Time and Cause: Essays Presented to Richard Taylor*, (pp. 109–136). Dordrecht: D. Reidel Publishing.

Shoemaker, S. (2011). Realization, powers and property identity. *The Monist*, *94*(1), 3–18.

Skyrms, B. (1984). Epr: Lessons for metaphysics. *Midwest Studies In Philosophy*, *9*(1), 245–255.
URL http://dx.doi.org/10.1111/j.1475-4975.1984.tb00062.x

Sutherland, R. (1983). Bell's theorem and backwards-in-time causality. *International Journal of Theoretical Physics*, *22*(4), 377–384.
URL http://dx.doi.org/10.1007/BF02082904

van Fraassen, B. C. (1982). The charybdis of realism: Epistemological implications of bell's inequality. *Synthese*, *52*(1), 25–38.

Wharton, K. B. (2007). Time-symmetric quantum mechanics. *Foundations of Physics*, *37*(1), 159–168.
URL http://dx.doi.org/10.1007/s10701-006-9089-1