Retrocausality at no extra cost

Peter W. Evans

November 27, 2014

Abstract

One obstacle faced by proposals of retrocausal influences in quantum mechanics is the perceived high conceptual cost of making such a proposal. I assemble here a metaphysical picture consistent with the possibility of retrocausality and not precluded by the known physical structure of our reality. This picture employs two relatively well-established positions—the block universe model of time and the interventionist account of causation—and requires the dismantling of our ordinary asymmetric causal intuition and our ordinary intuition about epistemic access to the past. The picture is then built upon an existing model of agent deliberation that permits us to strike a harmony between our causal intuitions and the fixity of the block universe view. I conclude that given the right mix of these reasonable metaphysical and epistemological ingredients there is no conceptual cost to such a retrocausal picture of quantum mechanics.

Key words: Retrocausality, Temporal symmetry, Interventionism, Quantum mechanics, Bell's theorem

1 Introduction

It is no secret that by permitting retrocausal influences in quantum mechanics local hidden variables can be used to account for the violation of Bell's inequality.[1] But can we be sure that such a retrocausal picture is *metaphysically tenable*? The metaphysics of retrocausality is often broached in the philosophical literature in and around discussions of time travel and causal paradoxes and there seems to be a general sentiment that there is nothing *manifestly self-contradictory* about the idea, strange though it may seem at first. There is, however, a significant challenge from the philosophy of physics literature: Maudlin (2002) claims that retrocausality is fundamentally at odds with the "metaphysical picture of the

The final publication is available at Springer via http://dx.doi.org/10.1007/s11229-014-0605-0.

[1] See, for instance, Argaman (2010), Costa de Beauregard (1953, 1976, 1977), Cramer (1980, 1986), Hokkyo (1988), Miller (1996, 1997), Price (1984, 1994, 1996, 1997, 2001, 2008, 2012), Price and Wharton (2013), Rietdijk (1978), Sutherland (1983, 1998, 2008), Wharton (2007, 2010, 2013a,b) and Wharton, Miller and Price (2011).

past generating the future" and thus cannot be entertained as a metaphysical possibility in a reality such as ours. The plausibility of Maudlin's metaphysical picture will not be of concern to us here.[2] The purpose of this paper is to counterbalance Maudlin's picture with a carefully considered metaphysical alternative that coheres with the possibility of retrocausality and that is not precluded by the known physical structure of our reality.[3]

This project does not introduce any explicitly new metaphysical ideas; on the contrary, the picture developed here is a conglomeration of developed material from various contexts that is merely collected together under the one roof. The goal here is to combine these ideas into a single coherent picture which will assist us in forestalling some perceived metaphysical problems with retrocausality. I begin by setting out in Section 2 two relatively well-established positions that will serve as a solid conceptual foundation upon which to develop our metaphysical picture: the block universe model of time in Section 2.1 and Woodward's (2003) interventionist account of causation in Section 2.2. There are then two metaphysical intuitions that must be dismantled. The first is our ordinary asymmetric causal intuition: in Section 3.1 I describe Price's (1996) argument from temporal symmetry against the plausibility of extending our asymmetric causal intuitions to the microscopic realm. The second is our ordinary intuition about epistemic access to the past: in Section 3.2 I present Dummett's (1964) argument against the possibility of 'bilking' that clears a logical space for retrocausality at the expense of our intuition that our past is necessarily epistemically accessible independent of our own future actions. The claim here is that quantum mechanics is a theory that occupies this logical space. This then opens the way to build a symmetric picture of causation: in Section 4 I sequester Price's (2007) model of agent deliberation that permits us to strike a harmony between our causal intuitions, such as free will and unidirectional causation, and the picture derived from spacetime physics that future events are fixed within a deterministic and causally symmetric framework.

It is often claimed that the introduction of retrocausal influences into the interpretation of quantum theory is a higher conceptual cost to pay than the problems associated with the rejection of local hidden variables. One way to imagine the choice between these competing interpretational schema is in terms of *ideological economy*.[4] According to Quine (1951), the ideas that can be expressed in a theory comprise the ideology of the theory. The ideological economy of a theory is then a measure of the economy of primitive undefined statements employed to reproduce this ideology; fewer primitive statements imply a more economical theory. Thus, the argument often made against retrocausality is that its ideology is less economical than rejecting local hidden variables (and thus, perhaps, embracing action-at-a-distance in quantum mechanics). The goal of this paper

[2]Though Evans, Price and Wharton (2013) provide some discussion of this point.

[3]By this I mean the nomological or mathematical structure of our physical theories, rather than the ontology of the interpretation of that structure, which may or may not preclude the possibility of retrocausality.

[4]This terminology is derived from Quine's (1951) distinction between the *ontology* and the *ideology* of a theory.

is to show that introducing retrocausality incurs no ideological cost because the ingredients required to build a retrocausal quantum picture of reality are given to us for free by the metaphysical structure of our existing physical theories and the epistemological structure of our experiences.

2 Foundations

2.1 The block universe model of time

We begin building our metaphysical picture with a temporal model popular among many physicists and philosophers: the block universe model of time. Rather than modelling reality as a three dimensional space evolving under the passage of time, reality is envisaged according to this view as a four dimensional block of which time is a mere passive ingredient. In the philosophical literature the block universe view is thought of as characterised by two claims: the past, present and future are equally real; and there is no privileged instant nor objective flow of time through the block. The spatial and temporal relations between all events in the four dimensional block are thus on an equal footing and exist atemporally. This view thus forges a strong analogy between the conception of time and our ordinary conception of space. Just as there is nothing objective about labelling a particular position in space 'here' nor claiming the contents of 'here' to be more real than the contents of 'there', there is nothing objective about labelling a particular time 'now' whose contents can be thought of as any more real than the contents of any other position within the block.

The block universe model of time is consistent with a deterministic model of reality. In a deterministic physical model, specifying data along a single hypersurface of spacetime is sufficient to determine the events of spacetime past and future of the hypersurface. Similarly, the block universe view is committed to the reality of all the events of four dimensional spacetime.

2.2 The interventionist account of causation

The interventionist account of causation is introduced and defended by Woodward (2003). The essential ingredient in this account is the notion of *manipulability* or *control*: according to this account, we say that C is a cause of E just in case there is some possible intervention that can be carried out on C independently of the remainder of the system that will change E in some way or other, holding fixed all other properties of the system containing C and E (that is, the intervention excludes other causes of E that are independent of the intervention). Woodward's account is explicitly counterfactual in the sense that there need be only some *possible* intervention that can be made on C to bring about a change in E. The advantage of this account is that it can be employed to provide causal explanations without requiring that there exists a complete description of some spatiotemporal process connecting C and E. To understand this account of causation

more clearly, let us consider an illustrative example.

Imagine the ignition system of a car. It seems that we would want to say that the turning of the key in the ignition (event K) is the cause of the starting of the car's engine (event E). According to the interventionist account we can say that K is indeed the cause of E since it is possible to carry out an intervention on K, by not turning the key say, that will change E in some way or other, in this case the engine would simply not start, provided all the other elements contributing to the system were held fixed. We can in fact claim a causal connection here without explicitly spelling out the mechanism by which the turning of the key brought about the starting of the engine. However, this does not mean that we cannot spell out such a mechanism if we wished.

Consider the mechanical chain of events connecting the turning of the key to the starting of the engine: turning the key (event K) completes the circuit between the car's battery and the starter motor (event C) which then starts the starter motor spinning (event S); the spinning starter motor then turns over the drive shaft of the engine (event D) which starts the pistons drawing in and then combusting the fuel (event P); the combusting fuel powers the engine to start running (event E). We have a chain of events, $K \rightarrow C \rightarrow S \rightarrow D \rightarrow P \rightarrow E$, with a mechanical account of how each event brings about the next. However, the content of any causal claim about any two of these events according to the interventionist account of causation is not that there exists a mechanical connection between the events. The key to the interventionist account is to imagine that each of these events is a handle or variable that can be manipulated and controlled. Accordingly, what makes each event the cause of the next is the fact that there exists a functional dependency between the variables; that is, some possible intervention on a particular variable will (over a range of conditions) bring about a consistent change in the values of the variables further down the chain. If we were to intervene on the above system by replacing the battery with an old or faulty battery, the starter motor would fail to spin, thus changing the value of the variable associated with event S (on or off, say) from what it would have been had we not made the intervention.

There are two issues which arise from this account of causation that will be crucial to our characterisation of retrocausality later in this paper. Firstly, it will be beneficial for our purposes here to view the interventionist account of causation as a kind of genealogical account of how we, as agents, come to acquire the concept of causation in cases where we have no possibility of intervening in the world around us. To begin demonstrating how this might be the case, consider the way we might employ causal concepts to describe a situation in which it is impossible for us as humans to intervene on the system. The gravitational pull of the moon is responsible for the ebb and flow of the tides, and we would want to say that the moon *causes* the tides' ebb and flow. Even though it is implausible for us to actually manipulate and control this system, we can attribute our causal intuitions in this sort of case to an ability to extend our causal intuitions from cases in which we can manipulate and control. Through our knowledge of the gravitational interaction between

the moon and the tides, we can predict with confidence what the effect of some imagined (but perhaps physically impossible) intervention would be if we could in fact bring it about. It is this sort of knowledge which we usually gain by physical intervention and experimentation that allows us to make claims about the causal relations that exist within a system. Thus, it seems reasonable that we extend these causal notions to cases in which we do not in fact have the requisite ability to manipulate and control.

Secondly, it is interesting to note that the interventionist account of causation is not explicitly reliant on a particular temporal direction. The direction of causation is dictated by the nature of the functional dependencies between the relevant variables describing a system and the nature of the (actual or hypothetical) intervention we are considering. It is the epistemic relation that we hold with respect to the different variables involved in the intervention that align the direction of causation with the future temporal direction; we control (and thus usually know the significant preconditions of) the intervention but do not have epistemic access to the effect in the future independently of this control. We will take this up again in Section 4 below.

I mention these features of the interventionist account here to highlight the fact that an arguable consequence of this view is that our role as agents in the world can be seen as at the root of our concept of causation. While I leave further discussion until later, for now I simply wish tentatively to broach the outcome of this genealogical sketch that a being who interacted with the world differently to us as an agent would have a very different concept of causation to the one that we have.

With these metaphysical foundations in mind, let us now move on to dismantling two of our ordinary temporal intuitions.

3 Dismantling intuitions

3.1 Macroscopic intuitions, microscopic symmetry

A familiar intuition, indeed one that seems almost trivial, is that the properties of interacting systems are independent before they interact. This is built upon the observation of many instances where this apparent principle holds true. In macroscopic systems, where our physical descriptions are coarse-grained and statistical considerations are relevant, we take this principle for granted. However, Price (1996, 1997) asks the question whether we are justified in extrapolating this familiar macroscopic principle to considerations of microscopic systems, which are far more fine-grained and where statistical arguments are inapplicable. Let us consider Price's analysis.

Firstly, it seems that the origin of this principle is related to the asymmetry of thermodynamics. When systems evolve from states of disequilibrium (lower entropy) to states of equilibrium (higher entropy) it is because the initial conditions are special; namely, the initial conditions are low entropy. Thus, if we were to consider a macroscopic system evolving in the reverse temporal direction, it would look strange because it would appear

that highly correlated *incoming* influences were converging from disparate regions of space (imagine a pile of rubble 'un-collapsing' into a building) and these would be associated with a decrease in entropy. In such a case the violation of the principle that physical processes are uncorrelated before they interact would be a direct product of the violation of the second law of thermodynamics. Looking forwards in time again we can see that the temporal asymmetry which manifests itself in the correlations between *outgoing* influences is a result of special (low entropy) initial conditions and not a result of an inherent asymmetry within the laws of physics.

It appears to be assumed that this principle of outgoing correlations but incoming independence holds with respect to fine-grained physical systems. In such cases, unlike in the coarse-grained, statistical-level cases, such a temporally asymmetric perspective resists explanation in terms of boundary conditions. The boundary conditions explanation is usually applied in cases where the phenomena display the temporal asymmetry of entropy change. In a fine-grained system, such as that of two particles which come together, interact and then separate, there is no entropy gradient of the sort that arises from the coarse-grained description of macroscopic systems to indicate a temporal orientation to the interaction. The temporal reverse of the interaction would likewise lack an entropy gradient as in the ordinary temporal direction; this is a function of the temporal symmetry of the dynamical laws of the system. As Price (1996, 1997) argues, this undermines any imputed assumption that outgoing correlations exist in one direction and not in the other.

Furthermore, unlike in the coarse-grained case, there is no *observed* asymmetry in fine-grained systems that needs to be explained. Of course, we do indeed sometimes observe correlations between, say, disparate particles in microscopic systems, and we infer that the correlations arise from some past condition rather than some future condition, resulting in an asymmetric picture of any previous interaction between the particles. But as Frisch (forthcoming) argues, this inference is underpinned by an asymmetric assumption—the *initial randomness assumption*—that claims that the values of the variables contributing to the preconditions of the interaction are randomly distributed.[5] In other words, the initial randomness assumption makes use of the principle that the incoming influences are uncorrelated, the applicability of which is precisely what is at question for fine-grained systems. Thus, if we take Frisch's analysis seriously, then the asymmetry between the independence of incoming particles and the correlation of outgoing particles is not *directly* observed for fine-grained systems, but is simply inferred based upon our macroscopic intuitions.

Therefore we are left with a dichotomy between two physical principles at the fine-grained level: temporal symmetry in the dynamical laws of physical systems on the one hand and, on the other hand, the asymmetry of the independence of systems prior to interaction. Given this dichotomy, one could make quite a persuasive argument against the

[5]Frisch (2012) argues that this assumption is crucial for developing causal representations that play an important inferential and explanatory role in physics.

independence of microscopic systems prior to interaction purely on symmetry grounds. However, the point I wish to make here is not that the principle of incoming independence and outgoing correlations should be abandoned altogether for fine-grained systems. The goal of this paper concerns the ideological cost of a retrocausal quantum picture, an integral part of which is the possibility for incoming particles to be correlated with future measurement settings before interaction. In fine-grained systems where statistical arguments are inapplicable, like the quantum systems at the core of Bell's theorem, the macroscopic origins and asymmetric nature of the intuition underpinning the assumption that incoming particles are independent becomes starkly apparent. When this consideration is combined with the fact that quantum mechanics provides no direct observational evidence as to whether incoming systems are correlated with future measurement settings or not (since it is not possible to observe the unmeasured system without destroying the correlations—an issue to which we now turn), we see that there is sparse analytic grounds for ruling out temporally symmetric causation in quantum systems. Thus if we take these considerations seriously then the nature of the physics in this case does not preclude a picture of reality that coheres with the possibility of retrocausal influences.

3.2 The bilking argument

In our normal conception of causation, causes precede their effects. A causally symmetric viewpoint opens up the possibility that effects can precede their causes. This, however, immediately creates some potential conceptual difficulties. To demonstrate these difficulties, let us imagine a pair of events which we believe to be causally connected: a cause, C, and an effect, E. Let us further imagine that this connection is retrocausal; E occurs earlier in time than C. On first appearances it would then seem possible to devise an experiment which could confirm whether our belief in the causal connection is correct or not. Namely, once we had observed that E had occurred, we could then set about ensuring that C does not occur, thereby breaking any retrocausal connection that could have existed between them. If we were successful in doing this, then we would have *bilked* the effect of its cause. This is the bilking argument.

The bilking argument seems to drive one towards the claim that any belief an agent might hold in the positive correlation between event C and event E is simply false. If this were the case then the agent would have to give up any belief in retrocausal influences between C and E. Dummett (1964) disputes that giving up this belief is the only solution to the bilking argument. In exploring the terms under which a belief in retrocausality can be maintained, Dummett suggests that what the bilking argument actually shows is that a set of *three* conditions concerning the two events, and the agent's relationship to them, is incoherent. In any incoherent set of conditions, all three conditions cannot hold simultaneously. Thus, depending on which of these three conditions fails to hold, there may be scope for an agent to maintain a belief that the later cause retrocausally influences the earlier event. To motivate these conditions, let us consider Dummett's own example.

Dummett imagines a tribe to exist with the custom of sending young men on a lion hunt to prove their bravery. The men travel for two days, hunt for two days and spend two days on their return journey. Observers travel with the young men and report back to the chief of the tribe whether the men acquitted themselves with bravery or not. While the young men are away, the chief performs dances intended to cause the young men to act bravely. Significantly, he performs these dances for the whole six days, i.e. for two days during which the events that the dancing is supposed to influence have already taken place. The chief notices that on occasions when he dances, he subsequently learns that the young men had hunted bravely and, on occasions when he does not dance, he subsequently learns that the young men had hunted in a cowardly fashion. The chief thus observes there to be a positive correlation between his dancing and the young men's bravery and therefore maintains a belief in retrocausality.

Imagine further that we are to convince the chief that this practice of his were absurd. We arrange that the observers who had accompanied the hunt return early and report to the chief whether or not the young men had acted bravely. We then set a bilking challenge to the chief to dance if and only if the young men had *not* acted bravely. There are two possible outcomes of this challenge. If the chief accepts this challenge and dances then he must concede that his dancing does not ensure the bravery of the young men. Alternatively, imagine that the chief accepts the challenge and then discovers he is inexplicably unable to dance, i.e. his limbs will simply not move. Then the chief would have to admit that dancing is not an action which is within his power to perform. If this were to occur, however, it would then be fair to say that it is not the chief's dancing that causes the young men to be brave, rather it is the young men's bravery that makes possible his dancing. Thus, regardless of whether the chief dances or not, it seems that the chief must give up his belief in retrocausality.

It appears then that there are two incompatible conditions here concerning the chief's dancing: (i) there is a positive correlation between the chiefs dancing and the bravery of the young men; and (ii) dancing is within the power of the chief to perform. If the first condition is to hold, then the second condition must fail, and *vice versa*, as we have just seen. Dummett, however, suggests that an implicit third condition can be violated which allows both of these conditions to hold simultaneously and thus allows the chief to maintain his belief in retrocausality. To see this, let us first consider an agent who believes a certain action is effective in bringing about a *subsequent* event. Such an agent would believe the action to be the cause of the later effect. Dummett recognises that there is a connection between the foreknowledge the agent possesses about the subsequent event and the intention the agent has to perform the action. The agent only knows an event to occur in the future because they intend to bring it about by performing a certain action: the agent possesses knowledge in intention. This is in contrast to knowledge of the past which we can possess in more forms than merely in intention.

Let us then return to our example and imagine for the sake of argument that there

is a parallel between the knowledge that the chief can possess concerning the bravery of the young men and the case of foreknowledge described here, i.e. the chief only knows that the young men are brave due to his intention to dance. This would then make our bilking challenge inconclusive. Since we can no longer arrange that the observers report the behaviour of the young men to the chief, we can no longer force the occurrence of a negative correlation. If we further rule out that there are no inexplicable incidents when the chief is unable to dance, then we are left with the original situation whereby the chief merely observes a positive correlation between his dancing and the young men's bravery and the chief can thus maintain his belief in retrocausality. To arrive at this result we have had to jettison the following condition: (iii) the chief has epistemic access to the behaviour of the young men independently of his intention to dance. These three conditions form a set which is shown to be inconsistent by the bilking argument.

Let us state these conditions in the more general terms we encountered at the beginning of this section:

(i) There exists a positive correlation between an event C and an event E.

(ii) Event C is within the power of an agent to perform.

(iii) The agent has epistemic access to the occurrence of event E independently of any intention to bring it about.

An interesting point to notice at this stage is that these conditions do not specify in which order events C and E occur. If we consider why it is not the case that it is possible to bilk *future* effects of their causes, this is because condition (iii) fails to hold for future events. If knowledge about future events could be obtained independently of an agent's intention to perform certain actions, then it would be possible to bilk those future events of their causes; this would amount, in a way, to changing the events we already know to occur in the future. Since this sort of foreknowledge is not possible, we can consistently believe our actions to *bring about* the future. Conversely, if it were the case that some past event was known only through our intention to perform a certain action, then it would be consistent to believe our actions to *bring about* the past.

The conditions under which it is possible to maintain a belief in retrocausality are especially relevant to quantum mechanics. In fact, once we make a suitable specification of how condition (iii) can be violated, we find that there exists a strong parallel between the conditions which need to hold to justify a belief in bringing about the past and what we find to be the case in quantum mechanics. Following the prescription of Price (1996, p. 174), let us not suppose that a violation of condition (iii) entails that the relevant agent has *no* epistemic access to the relevant past events independently of any intention to bring them about, rather let us suppose that the means by which knowledge of these past events is gathered breaks the claimed correlation between the agent's action and those past events. We can state our new condition as follows:

(iv) The agent can gain epistemic access to the occurrence of event E independently of any intention to bring it about and without altering event E from what it would have been had no epistemic access been gained.

In the dancing chief example a violation of this condition would entail that every time the chief attempted to discover the behaviour of the young men he subsequently affected their behaviour to be different from what it would have been had he not attempted his discovery. In those cases where the chief makes no attempt to discover the behaviour of the young men, we are back to our original violation of condition (iii).

The nature of this weakened violation of condition (iii) should look familiar; it is just the sort of condition we would expect to hold if the system in question were a quantum system. To make this explicit, let us consider a simple quantum system as an example to see that this is the case. Imagine a quantum system is prepared to be in a state ψ at time t_0 and at time t_1 the system is to be measured; without loss of generality, let us say that we are making a spin measurement at t_1 (as we would do in a test of the Bell inequality). In the orthodox interpretation of quantum mechanics the wavefunction representing the system evolves according to the Schrödinger equation from t_0 until the time of measurement at t_1 wherein the wavefunction collapses to one or other of the eigenstates of the operator associated with the measurement.

Let us now imagine that we are an agent who believes in a retrocausal influence from the measurement setting at t_1 on the state of the system after preparation at t_0. To begin with, we cannot subscribe to the view that the wavefunction description is a complete description of the quantum state; due to the asymmetric nature of the measurement process and the Born rule, there is simply no correlation at all in orthodox quantum mechanics between the wavefunction after t_0 and the measurement setting at t_1. To make a claim that there is a correlation between the state of the quantum system after t_0 and the measurement setting at t_1, we simply cannot take the wavefunction to be a complete representation of the quantum system. Thus we must subscribe to a hidden variable account of quantum mechanics and interpret the wavefunction epistemically, as representative of the possible knowledge an observer can gain about the system. In doing so, we can stipulate through the subsequent quantum model that the hidden variables of the system after t_0 (corresponding to some spin states) are positively correlated to the measurement setting at t_1, i.e. that condition (i) holds. Furthermore, we will assume here that we have the power to perform any spin measurement we like at t_1, i.e. condition (ii) holds. Let us now consider whether condition (iv) can also hold; that is, whether it is possible to bilk our experiment.

Imagine how a bilking argument against our belief in a retrocausal influence might run. We begin with a claim that the values of the spin state of the system between t_0 and t_1 are correlated with the spin measurement setting we will freely choose at t_1. A potential bilker will attempt to somehow observe the actual spin state at some time $t_0 < t_b < t_1$ in such a way so as to establish a contradiction between the observation at t_b and the value

we would observe as a result of carrying out a measurement at t_1. However, according to quantum mechanics, as a result of such an observation there will be a new wavefunction description of the system after t_b that suggests the system is evolving from an eigenstate of the operator corresponding to the measured observable at t_b. Thus, in support of our belief in a retrocausal influence from the measurement setting at t_1 on the spin state of the system, we can claim that any correlation between the spin state after preparation at t_0 and the measurement setting at t_1 no longer obtains. The new wavefunction no longer represents the evolution of the spin state between t_0 and t_1, as the bilker's observation has destroyed any putative correlation between the state and the measurement setting at t_1; the new wavefunction is now tracking the spin state between t_b and t_1 and we can retrodict that the spin state between t_0 and t_b is correlated with the measurement settings chosen by the bilker at t_b. Thus the hidden variables characterising the system after the bilker's observation will not be what they would have been had the bilker not made that observation; condition (iv) will consequently be violated.

Therefore, the very nature of quantum mechanics ensures that any claimed positive correlation between the future measurement settings and the hidden variables characterising a quantum system cannot possibly be bilked of their causes because condition (iv) is perennially violated. Moreover, we can stipulate further that the intervening observation of the system by the bilker establishes a new correlation between the measurement settings of the observation at t_b and the hidden variables of the system between t_0 and t_b (and, by the same logic, with the confidence that this correlation cannot be bilked either). The significant detail of this picture is that (so long as we subscribe to the epistemic interpretation of the wavefunction) we lack epistemic access to the "hidden" variables of the system and we lack this access *in principle* as a result of the structure of quantum mechanics. (In the terminology of the next section, the hidden variables are UNKNOWABLES.)

Thus, on a certain interpretation of the wavefunction formalism in quantum mechanics, the hypothesis of retrocausality in quantum systems remains intact in the face of the bilking argument. In fact, according to Dummett's analysis of the bilking argument, quantum mechanics has exactly the sort of dynamics we would expect of a retrocausal physical theory; the *counterintuitive* nature of backwards-in-time causality can hardly be seen as a disadvantage here. The fundamental structure of quantum mechanics does not preclude a metaphysical picture that allows the possibility of retrocausal influences.[6]

4 Keeping up appearances

We now have a better idea of the sorts of limitations that constrain the picture of reality that allows the possibility of retrocausal influences. We are now in a position to use these constraints, along with the causal and spatiotemporal structures we have taken to be most reasonable, to build a picture of what retrocausality actually involves. At the centre of

[6]Nor are hidden variables of this sort precluded by any of the quantum no-go theorems concerning hidden variables, as these theorems assume that there is no retrocausality.

this discussion will be the role that we play as agents as we interact with, and participate in, the world.

Let us start first and foremost with two conceptions of influence that are commonly conflated when talking about the future: the view that we *change* events and the view that we *affect* events. Consider a claim like the following: by deciding to catch the bus, I changed my day from one in which I was late for work, to one in which I was early. Regardless of one's model of time, there is an inconsistency in thinking that we change events through our actions. For an event to 'change', the event must have been a particular way in the first place. If we were partial to a dynamic view of time in which the future were *unreal*, it would make no sense to think of a future event as being any particular way before it is actual; there is simply no event that is my tardiness which can be changed before I am in fact late. However, we have explicitly signalled our intention to employ the block universe model of time and in such a model we *can* speak of future events as being real and thus it might be possible for an event, one might say, to be a particular way *ab initio*. We might say that my tardiness was an event and that this event changed into my punctuality. But we must be careful here, because if a future event is real, it is in some sense already out there in the four dimensional block. If we *change* it at some point prior to it being a present event for us, we are left with the rather strange question: why was it as it was before we changed it? Why did the four dimensional block contain an event which was my tardiness, which then changed at some point into my punctuality? With respect to the block universe view this question does not make any sense.

Let us take stock and see if we can clarify the above claim. We might do this by saying something like the following: when we say that we change a future event, we mean that we change it from being something that it *could* have been, say my tardiness, to something that it now actually *is*, say my punctuality. Expressing what we mean by change in counterfactual terms lets us sidestep the problems we encountered with the reality of the events under question. However, the notion we have ended up with by doing so has a significant causal ring to it (recall our characterisation of causation in terms of interventions); this is in fact just what we mean when we use the word 'affect'. I affect my day to be a day in which I am early for work, rather than a day in which I am late. I play a particular role in bringing about the future event and it is wrong to think that I change it from something that it already was. As long as we commit ourselves to the block universe view in which all events in the past, present and future are equally real, then we must think of influence in the 'affect' sense. Furthermore, we can now see that this argument is as much relevant to past events as it is relevant to future events. Under no circumstances does it make sense to change the past in any way, since one cannot change something that is already an actual event. Retrocausality is then not about changing the past, rather retrocausality is about affecting the past: playing a role in bringing about a past event.

This analysis is beginning to push us into a position about determinism and the nature

of the block universe that may seem highly undesirable; namely, that we have no freedom in choosing our own actions. If we cannot change the future in just the same way that we cannot change the past, and if affectation is merely bringing about an event that in some sense already exists, then it would seem that we are mere spectators of our reality in a rather uninteresting sense. Fortunately, we are not pushed into this position by adopting typically block universe notions as above. Moreover, coming to grips with why this is the case will tie together many of the issues with which we have so far dealt and it will give us our first glimpse at the metaphysical picture of reality that allows for retrocausal influences.

The solution to this seeming incompatibility between the conception of reality as a block universe and our ability as agents to control and manipulate our surroundings lies in thinking of causation as a perspectival notion. According to Price (2007), evidence suggests that causation is indeed a perspectival notion; we have already been introduced to the idea when we were considering the interventionist account of causation above. The tentative outcome that I flagged of what we called a kind of genealogical account of causation in terms of intervention was that a being who interacted with the world differently to how we interact with the world as agents (i.e. has a different perspective of the world) would have a different concept of causation to the one that we have. Let us consider how we can use this to help us find some sort of compatibility between the block universe view and our causal intuitions.

The essential point to solving this problem is to realise that considering the block universe 'from the outside' is availing oneself of a very different perspective of the world to the one which we have while we are inhabiting some spatiotemporal region. The important difference between the two viewpoints is that there is a discrepancy between the parts of the spacetime block that are epistemically accessible from each perspective. The spatiotemporally constrained perspective by which we are bound permits us only limited epistemic accessibility to other spatiotemporal regions. This is significant because it is as spatiotemporally bound agents that we have evolved and it seems reasonable to suggest that we are in possession of a concept of causation that reflects this very fact. Once we imagine ourselves to be omniscient beings that have epistemic access to the whole spatiotemporal block, as we have done in the above analysis of change and affect, it should not come as a surprise that our causal intuitions get confused when we attempt to consider how a spatiotemporally bound agent can deliberate about whether or not to affect a particular event that is already determined from our imagined omniscient perspective. The solution that I am pushing towards here is that it is because we do not *know* which events are determined to occur that we can deliberate, and therefore be causal agents, at all. To reach this conclusion we sequester one final analysis concerning the relationship between deliberation and epistemic accessibility, and the role this plays in our concept of causation.

Price (2007, p.20) sets out "an abstract characterisation of the structural, or func-

tional architecture, of deliberation" with a view to separating out the intrinsic features of deliberation itself from those aspects of deliberation that are a function of our perspective as spatiotemporally bound agents. To begin with, a deliberator must be deliberating over whether to bring about some particular occurrence out of a range of possible occurrences. Following Price, we will call the set of events of which this range consists the OPTIONS that the deliberator is considering. The set OPTIONS can be thought of as consisting of two subsets: all those occurrences over which the deliberator has immediate control, the DIRECT OPTIONS, and all those occurrences that can be brought about indirectly via the DIRECT OPTIONS, the INDIRECT OPTIONS. All other events that are not under consideration during the deliberation we will call the FIXTURES. An integral subset of the FIXTURES is the set of events that the deliberator already knows, or are in principle knowable, at the time of deliberation which we will call the KNOWABLES. The KNOWABLES must be a subset of the FIXTURES since if these events are knowable to the deliberator at the time of deliberation, then they cannot be under consideration to be brought about and thus cannot be part of the set OPTIONS. For this reason, all the events in OPTIONS must fall into the set we will call UNKNOWABLES. Thus a deliberator makes two dichotomous distinctions: the distinction between FIXTURES and OPTIONS; and the distinction between KNOWABLES and UNKNOWABLES. The set KNOWABLES is a subset of FIXTURES and the set OPTIONS is a subset of UNKNOWABLES. Let us now consider how spatiotemporally bound deliberators such as ourselves might map these distinctions onto the past and the future.

Considering the future first, we are going to want to say that much of the future belongs to the set FIXTURES. This is largely due to the finite nature of deliberation: since we do not deliberate about bringing about the whole future all at once, there are then many future occurrences that we take as part of the fixed background during the deliberative process. It also seems as given that the set DIRECT OPTIONS must also be comprised of future events. We can attribute this to the fact that we are temporally constrained agents of a particular sort; the set DIRECT OPTIONS consists of our immediate actions and we simply cannot deliberate about whether to bring about our past actions, only our future actions. Further to this, we might want to say that the set INDIRECT OPTIONS also is comprised exclusively of future events, but this would be so only if we were committed to classifying all past events as belonging to the set FIXTURES. Ordinarily, this is exactly how we consider past events: as fixed. This is for the most part a function of the fact that we consider the past as knowable in principle, and as we have seen above, the set KNOWABLES is a subset of the set FIXTURES. But is it the case that our spatiotemporally bound perspective commits us to the past being fixed?

If such a commitment is indeed a function of the fact that we consider the past as knowable in principle, then it would seem that the possibility of the past being unknowable in principle would purge us of this commitment. Recall that this is exactly the condition we found to be suitable to avoid the bilking argument in the above analysis of Dummett:

an agent is immune to having a belief in a particular retrocausal correlation bilked if the past effect in question is epistemically inaccessible to the agent at the time of the causal action. In the language of our current analysis, if some past event belongs to the set UNKNOWABLES then it does not *necessarily* belong to the set FIXTURES, and thus an agent is not precluded from counting the event as belonging to the set INDIRECT OPTIONS.

To see what this entails, consider the result from the last section that the structure of quantum mechanics ensures that certain facts concerning putative hidden variables are immune to the bilking argument. Thus, in the right circumstances, there is information about the past of some quantum systems that is epistemically inaccessible *in principle*. This implies, as we noted above, that hidden variables should belong to the set UNKNOWABLES. But now we see that it is not implausible that hidden variables (along with perhaps other microscopic degrees of freedom that are epistemically inaccessible) belong to the set INDIRECT OPTIONS. If this is the case then it is a live possibility that the set INDIRECT OPTIONS contains some events which we would take to be past. Since the set INDIRECT OPTIONS are just those events we take to be in our power to bring about, the architecture of deliberation does not rule out the possibility of bringing about past events.

It is interesting to consider this result in light some of the analysis of this paper. There is subtle connection between the KNOWABLES/UNKNOWABLES distinction and the macroscopic/microscopic distinction. If we imagine a case wherein the value of some hidden variable—say, a spin state—has been observed through some measurement procedure, we would take such microscopic degrees of freedom to then be irreversibly coupled to a macroscopic system that registers the measured value. This renders the observed spin state part of the set KNOWABLES, and thus it would no longer be possible for these variables to belong to the set INDIRECT OPTIONS. Indeed, one might argue that the observation of a microscopic system simply consists in coupling to a macroscopic system (in many cases that macroscopic system would just be us), and this would demonstrate the connection between the KNOWABLES/UNKNOWABLES distinction and the macroscopic/microscopic distinction. Thus, any claim to a retrocausal influence in a physical system must be confined to the microscopic degrees of freedom, such as the hidden variables in a quantum system, since it is only those events in the set UNKNOWABLES that meet the criteria set out in this analysis.

This schematic of where retrocausality fits in to the structure of deliberation highlights an important feature of a metaphysical picture that allows retrocausal influences: that agents within such a reality will always deliberate towards the future, i.e. the set DIRECT OPTIONS will always be comprised of future events. Thus retrocausality is not deliberation towards the past, or in other words, it is not our normally directed causation in the reverse temporal direction.

The way that any particular agent divides the set of all events into FIXTURES and OPTIONS, KNOWABLES and UNKNOWABLES and past and future will depend completely upon the agent's spatiotemporal perspective. For spatiotemporally constrained agents

such as ourselves, there is a specific recipe for how these distinctions are made which is a function of the way we have evolved from within the spacetime block. If we imagine ourselves as omniscient beings who are observing the events in the spacetime block 'from the outside', all the events will be in the set KNOWABLES and thus in the set FIXTURES. This is how we can imagine the spacetime block to be entirely determined without having this intuition be in conflict with our usual sense of free choice in the deliberative process; these are vastly different perspectives and causality is perspectival. It is the extent of our ignorance, of both the future and of the complete set of prior causes of our actions, that creates the illusion, so to speak, of free choice. This is where we then strike a harmony between our causal intuitions, such as deliberation, and the intuition that future events are fixed within a deterministic framework. The crucial element is to realise that we, as spatiotemporally bound agents, are constrained in our epistemic access to the events in spacetime.

5 Conclusion

This then is the package of metaphysical ideas that combine to give a picture that is consistent with the possibility of retrocausality. We begin with two established metaphysical foundations in the block universe model of time and the interventionist account of causation. We then remove two potential obstacles originating in our ordinary temporal intuitions: we realise that we have no evidence to suggest our macroscopic asymmetric causal intuitions can be extrapolated to the microscopic realm and we realise that we do not necessarily have epistemic access to the past independent of our own future actions. With these obstacles gone, the emerging picture of a temporally and causally symmetric reality viewed from an epistemically limited vantage point concords well with the possibility of retrocausality. A significant aspect of this assembly of ideas is that none of the included elements are precluded by the known physical structure of our reality. Indeed, if anything, these elements are supported by the structure of at least one of our best physical theories: quantum mechanics.

I began this discussion by claiming that a retrocausal quantum picture of reality required no new 'ideology' in order to be plausible. However, one might argue that, while perhaps no new ideology has been introduced, certain metaphysical issues, in particular our notion of free choice, have become harder to reconcile with orthodox positions on the matter. With respect to the analysis of the previous section, and true to my claim that the required ingredients are already available to us, I point out that all I have espoused here is a *compatibilist* notion of free will. Indeed, anyone interested in supporting an eternalist position with respect to the ontology of time—wherein the past, present and future are equally real—must deal with this very issue. Of course, this is not a treatise nor defence of the compatibility of free will with, say, moral responsibility, but at the very least compatibilism is a major player in this debate and could be considered a more than plausible solution to the problem of free will. I would certainly consider this element of

the picture presented here to fit within an orthodox camp on this matter.

But further questions do arise concerning the metaphysical consequences of a retrocausal quantum picture of reality. For instance, what constraint determines the structure of our epistemic access to the past and future? While we can point to the asymmetric boundary conditions of the universe and the resulting entropy gradient as the beginnings of an answer, this issue indeed requires further attention (and is presumably an issue that receives a more straightforward response in Maudlin's picture of the past generating the future). More significantly, what sort of model of quantum mechanics is going to be able to support this metaphysical picture? On this matter, I wish to mention two promising areas of research that are arguably moving in the right direction to providing such support. Firstly, Spekkens (2007) outlines a toy theory that "reproduces in detail a large number of phenomena that are typically taken to be characteristically quantum"; whilst Spekkens' project does not concern retrocausal quantum mechanics, it is based on a constraint that an observer has epistemic access to only half of the information concerning reality. If this sort of epistemic constraint is found to underlie quantum mechanics, then the picture presented in this paper may be a useful guide to further understanding on this issue. Secondly, Wharton (2007, 2010, 2013a) pursues an explicitly retrocausal theory of quantum mechanics based on the determination of physical behaviour by both initial and final boundary conditions. Since we ordinarily lack epistemic access to final boundary conditions, this again creates the sort of conditions integral to the retrocausal quantum picture of reality I present here. While these works do not present a fully developed theory or interpretation of quantum mechanics, they do suggest a way forward for the retrocausal program.

While Maudlin is clearly correct in noticing that retrocausality is fundamentally at odds with the metaphysical picture of the past generating the future, this by no means renders retrocausality metaphysically untenable. Given the right mix of some reasonable metaphysical and epistemological ingredients, an alternative metaphysical picture arises that is consistent with the possibility of retrocausality. Moreover, the ideological cost of these ingredients cannot outweigh the interpretational problems associated with the rejection of local hidden variables, such us the acceptance of action-at-a-distance in quantum mechanics, simply for the fact that we were given all these ingredients for free by the metaphysical structure of our existing physical theories and the epistemological structure of our experiences.

Acknowledgements

I wish to thank Sam Baron, John Cusbert, Matt Farr, Huw Price, Mikey Slezak and two anonymous referees for *Synthese* for helpful discussions and comments. This research has been supported partly by the Australian Research Council and the *Scholarship for Research on Foundations of Quantum Mechanics* at the Centre for Time, University of Sydney and partly by the Templeton World Charity Foundation grant: *The causal power of information in a quantum world* at the University of Queensland.

References

Argaman, N. (2010). Bell's theorem and the causal arrow of time. *Am. J. Phys.* **78**: 1007–1013. doi:10.1119/1.3456564. arXiv:0807.2041 [quant-ph].

Costa de Beauregard, O. (1953). Méchanique Quantique. *Comptes Rendus de l'Académie des Sciences* **T236**: 1632–1634.

——— (1976). Time Symmetry and Interpretation of Quantum Mechanics. *Found. Phys.* **6**: 539–559. doi:10.1007/BF00715107.

——— (1977). Time symmetry and the Einstein paradox. *Il Nuovo Cimento* **42**: 41–63. doi:10.1007/BF02906749.

Cramer, J. G. (1980). Generalized absorber theory and the Einstein-Podolsky-Rosen paradox. *Phys. Rev. D* **22**: 362–676. doi:10.1103/PhysRevD.22.362.

——— (1986). The transactional interpretation of quantum mechanics. *Rev. Mod. Phys.* **58**: 647–687. doi:10.1103/RevModPhys.58.647.

Dummett, M. (1964). Bringing About the Past. *The Philosophical Review* **73**(3): 338–359.

Evans, P. W., Price, H. and Wharton, K. B. (2013). New Slant on the EPR-Bell Experiment. *Brit. J. Phil. Sci.* **64**: 297–324. doi:10.1093/bjps/axr052. arXiv:1001.5057v3 [quant-ph].

Frisch, M. (2012). No place for causes? Causal skepticism in physics. *Euro. J. Phil. Sci.* **2**(3): 313–336. doi:10.1007/s13194-011-0044-4.

——— (forthcoming). Causes, Randomness, and the Past Hypothesis. *In* B. Loewer, E. Winsberg and B. Weslake (eds.), *Time's Arrows and The Probability Structure of The World*, MIT Press, Cambridge, Massachusetts.

Hokkyo, N. (1988). Variational formulation of transactional and related interpretations of quantum mechanics. *Found. Phys. Lett.* **1**: 293–299. doi:10.1007/BF00690070.

Maudlin, T. (2002). *Quantum Non-Locality and Relativity.* Blackwell Publishing, Oxford.

Miller, D. J. (1996). Realism and time symmetry in quantum mechanics. *Phys. Lett. A* **222**: 31–36. doi:10.1016/0375-9601(96)00620-2.

——— (1997). Conditional probabilities in quantum mechanics from time-symmetric formulation. *Il Nuovo Cimento* **112B**: 1577–1592.

Price, H. (1984). The philosophy and physics of affecting the past. *Synthese* **61**: 299–324. doi:10.1007/BF00485056.

——— (1994). A Neglected Route to Realism about Quantum Mechanics. *Mind* **103**: 303–336. doi:10.1093/mind/103.411.303.

——— (1996). *Time's Arrow and Archimedes' Point.* Oxford University Press, New York.

——— (1997). Time Symmetry in Microphysics. *Phil. Sci.* **64**: 235–244. arXiv:quant-ph/9610036v1.

——— (2001). Backwards causation, hidden variables, and the meaning of completeness. *Pramana J. Phys.* **56**: 199–209. doi:10.1007/s12043-001-0117-6.

——— (2007). Causal Perspectivalism. *In* H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press, Oxford, chapter 10, pp. 250–292.

——— (2008). Toy models for retrocausality. *Stud. Hist. Phil. Mod. Phys.* **39**: 752–776. doi:10.1016/j.shpsb.2008.05.006.

——— (2012). Does Time-Symmetry Imply Retrocausality: How the Quantum World Says "Maybe". *Stud. Hist. Phil. Mod. Phys.* **43**: 75–83. doi:arXiv:1002.0906 [quant-ph].

Price, H. and Wharton, K. B. (2013). Dispelling the Quantum Spooks—a Clue that Einstein Missed? arXiv:1307.7744 [physics.hist-ph].

Quine, W. V. O. (1951). Ontology and Ideology. *Phil. Stud.* **2**: 11–15. doi:10.1007/BF02198233.

Rietdijk, C. W. (1978). Proof of a retroactive influence. *Found. Phys.* **8**: 615–628. doi:10.1007/BF00717585.

Spekkens, R. W. (2007). Evidence for the epistemic view of quantum states: A toy theory. *Phys. Rev. A* **75**: 032110. doi:10.1103/PhysRevA.75.032110.

Sutherland, R. I. (1983). Bell's theorem and backwards-in-time causality. *Int. J. Theor. Phys.* **22**: 377–384. doi:10.1007/BF02082904.

——— (1998). Density Formalism for Quantum Theory. *Found. Phys.* **28**: 1157–1190. doi:10.1023/A:1018850120826.

——— (2008). Causally symmetric Bohm model. *Stud. Hist. Phil. Mod. Phys.* **39**: 782–805. doi:10.1016/j.shpsb.2008.04.004.

Wharton, K. B. (2007). Time-Symmetric Quantum Mechanics. *Found. Phys.* **37**: 159–168. doi:10.1007/s10701-006-9089-1.

——— (2010). A Novel Interpretation of the Klein-Gordon Equation. *Found. Phys.* **40**: 313–332. doi:10.1007/s10701-009-9398-2.

——— (2013a). Lagrangian-Only Quantum Theory. arXiv:1301.7012 [quant-ph].

——— (2013b). The Universe is not a Computer. *New Scientist* **217**: 30–31. doi:10.1016/S0262-4079(13)60354-1. arXiv:1211.7081 [quant-ph].

Wharton, K. B., Miller, D. J. and Price, H. (2011). Action Duality: A Constructive Principle for Quantum Foundations. *Symmetry* **3**: 524–540. doi:10.3390/sym3030524. arXiv:1103.2492 [quant-ph].

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.