Synthese
DOI 10.1007/s11229-012-0195-7

# How to evaluate counterfactuals in the quantum world

**Tomasz Bigaj**

Received: 15 August 2012 / Accepted: 15 August 2012


**Abstract** In the article I discuss possible amendments and corrections to Lewis's semantics for counterfactuals that are necessary in order to account for the indeterministic and non-local character of the quantum world. I argue that Lewis's criteria of similarity between possible worlds produce incorrect valuations for alternate-outcome counterfactuals in the EPR case. Later I discuss an alternative semantics which rejects the notion of miraculous events and relies entirely on the comparison of the agreement with respect to individual facts. However, a controversy exists whether to include future indeterministic events in the criteria of similarity. J. Bennett has suggested that an indeterministic event count toward similarity only if it is a result of the same causal chain as in the actual world. I claim that a much better agreement with the demands of the quantum-mechanical indeterminism can be achieved when we stipulate that possible worlds which differ only with respect to indeterministic facts that take place after the antecedent-event should always be treated as equally similar to the actual world. In the article I analyze and dismiss some common-sense counterexamples to this claim. Finally, I critically evaluate Bennett's proposal regarding the truth-conditions for true-antecedent counterfactuals.

**Keywords** Counterfactuals · Indeterminism · Quantum mechanics · True-antecedent counterfactuals

## 1 Lewis's semantics for counterfactuals

Among the most prominent achievements of the twentieth-century formal semantics one should unquestionably count the logical analysis of counterfactual conditionals

T. Bigaj (✉)
Institute of Philosophy, University of Warsaw, Warsaw, Poland
e-mail: t.f.bigaj@uw.edu.pl

Published online: 05 October 2012

(subjunctive conditionals) given by David Lewis.[1] Lewis proposes to interpret counterfactuals as "variably strict conditionals". Strict conditionals, as it is well known, are characterized as conditionals that are true *necessarily*. Within the standard possible-world semantics this characteristics can be expressed as stipulating that for a strict conditional to be true, its consequent has to be true in all possible worlds (that are accessible from the actual world) in which the antecedent is true. But Lewis observes that this requirement is too strong for counterfactual conditionals. For a statement "If John jumped from the thirteenth floor of the Empire State Building, he would be killed" to be true, the consequent "John was killed" does not have to be true in all possible worlds in which the antecedent "John jumped from the thirteenth floor" is true. After all, there are perfectly imaginable worlds in which humans learned how to fly, or—perhaps more plausibly—in which every building in New York City more than ten story high has a safety net around it.

As Robert Stalnaker first noted in his 1968 article, for a counterfactual to be true, its consequent has to be true in all antecedent-worlds which are reasonably similar to the actual world as we know it. Lewis picked up on this idea and introduced the notion of a *comparative similarity* (*closeness*) *between possible worlds*. In his semantics the set of possible worlds is endowed with a binary relation $\leq$ whose intuitive sense is such that for two worlds $w_1$ and $w_2$, $w_1 \leq w_2$ iff $w_1$ is at least as similar to the actual world as $w_2$ is. Formally, this relation is supposed to fulfill three requirements: the requirement of transitivity, of strong connectedness, and of minimality with respect to the actual world $w_0$. With the help of the relation of comparative similarity $\leq$, it is possible to formulate the celebrated truth-condition for the counterfactual (the symbol "$\Box\rightarrow$" represents the counterfactual connective):

(L) $P\Box\rightarrow Q$ is non-vacuously true at $w_0$ iff there is a $P$-world $w_1$ in which $Q$ is true, and there is no $P$-world $w_2$ such that $w_2 \leq w_1$ and not-$Q$ is true in $w_2$.

Understandably, the purely formal characteristic of the similarity relation $\leq$ given by Lewis is not sufficient for the task of evaluating particular counterfactual conditionals that can be encountered in the natural language, as well as in the language of science. To do this we need to put some flesh on the bones of Lewis's formal system. Lewis himself has proposed a particular way of comparing possible worlds with respect to their relative similarity to the actual world, of which we shall talk more later. The main purpose of this paper is to assess informal criteria of inter-world similarity from the perspective of their applicability to the analysis of quantum counterfactuals, i.e. counterfactual conditionals that appear in the context of quantum theory. I shall argue that Lewis's similarity relation is in need of serious corrections due to some peculiarities of the quantum-mechanical description, such as indeterminism and nonlocality. However, I believe that the amendments and corrections that shall be proposed here are not uniquely tied to the quantum case, and that they can actually help better analyze some natural-language counterfactuals too. Thus the discussion provided in the paper should be of interest not only to the philosophers of quantum mechanics, but

[1] His most authoritative work on this subject is the book *Counterfactuals* (Lewis 1973).

to the philosophers of language as well.[2] But first we have to say a few words about why counterfactuals are needed in the foundational analysis of quantum mechanics in the first place.

## 2 Counterfactuals and quantum mechanics

The last two decades have witnessed a noticeable rise in the number of publications in which counterfactual logic has been applied to the analysis of some interpretational issues in quantum mechanics. While it is still uncertain whether this new approach can lead to new and groundbreaking results (and we have to remember that many initially promising approaches to the foundation of quantum mechanics subsequently turned out to be blind alleys), it is clear that more and more philosophers and physicists turn their attention to this conceptual tool.[3] Counterfactual logic is mostly (but not exclusively) used in the context of the so-called entangled quantum systems, i.e. systems consisting of two or more objects, whose parameters display peculiar correlations, entirely predicted by the quantum theory and confirmed in various experiments. Due to these correlations, it may be for example possible to disclose the value of a given parameter for one particle, based on the result of an experiment performed on the other particle. This fact creates a conceptual problem for the orthodox interpretation of quantum mechanics, according to which in most typical cases a value of a parameter does not exist before it is directly measured. The aforementioned problem has been spelled out in the most dramatic fashion in the famous and celebrated Einstein–Podolsky–Rosen (EPR) argument against quantum orthodoxy (cf. Einstein et al. 1935), whose foundations were laid down by Niels Bohr and Werner Heisenberg.

There are three basic reasons why the application of counterfactuals in the analysis of such issues can be expedient. Firstly, quantum mechanics places a severe restriction on the number of parameters whose values can be determined at a given time. Among several parameters that may separately characterize a given quantum particle, we can select only a handful of "compatible" quantities such that a measurement revealing the value for one of them does not "destroy" the values of the other ones. Incompatible parameters, like position and momentum, or various components of spin, cannot have their values jointly determined. Hence, if we decide to measure a particle's momentum, its position becomes undefined, and vice versa. And yet sometimes we want to talk about alternative measurement settings, in which a parameter incompatible with the one actually selected would have been put to an experimental test. For example, the aforementioned EPR argument considers a situation in which a given parameter has

[2] This feature of the paper shall be reinforced by discussions of the recent interpretations of some natural-language counterfactuals proposed by Jonathan Bennett in his 2003 book.

[3] The master champion of this new counterfactual approach is Henry P. Stapp from Lawrence Berkeley National Laboratory (cf. Stapp 1971, 1989, 1997, 1998, 2001; Bedford and Stapp 1995). His main objective has been to use the counterfactual semantics in order to strengthen the so-called Bell theorem. It needs to be mentioned, however, that all his attempts have been subjected to a severe criticism, among others in Redhead (1987), Clifton et al. (1990), Mermin (1998), Shimony and Stein (2001), and Bigaj (2006, 2007). Examples of other uses of counterfactuals in quantum mechanics, not limited to the reinterpretation of the Bell theorem, can be found in Ghirardi and Grassi (1994), Griffiths (1999, 2001), and Vaidmann (1999).

been selected to be measured for one of the two entangled particles, but then it moves on to the analysis of what would have happened, had we decided to measure an alternative parameter that is not compatible with the previous one. To speak meaningfully of such situations, one has to have a firm grasp of the meaning of the counterfactual phrases.

The EPR argument uses as its crucial premise the so-called principle of locality, which also constitutes one of the fundamentals of Albert Einstein's scientific *Weltanschaung*. Loosely speaking, this principle prohibits the existence of physical interactions that would propagate instantaneously, without any delay (so-called action at a distance). In other words, an event $e$ occurring at one location cannot be a cause of any distant change that would happen simultaneously with $e$.[4] Yet these formulations of the locality condition are hardly satisfactory, for they contain notoriously ambiguous and open to interpretations notions like "interaction" or "cause". For that reason it may be useful to employ once again counterfactual conditionals in order to spell out precisely the intended meaning of the locality requirement. One such possible explication may be as follows: we can say that the locality condition is satisfied, iff for any possible localized event $e$ it is true that if $e$ occurred, the entire part of space–time that is space-like separated from $e$ would remain exactly the same as in the actual world. This expression (and other alternatives as well) can be given a precise meaning with the help of a formal semantics of the counterfactual.

Finally, there is yet another aspect of the quantum-mechanical description beside the two given above that may require a counterfactual reformulation. As it is well known, Einstein supplements his EPR argument with the so-called criterion of physical reality: "if, without in any way disturbing a system, we can predict with certainty (·) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity". But how are we supposed to understand the phrase "there exists an element of physical reality"? Typically, it is assumed that it refers to the objective fact of possessing a given measurable property by a system. In other words, if we can infer—without actually performing a direct measurement—that the value of a given parameter characterizing a particle should be such-and-such, then the particle can be assumed to objectively possess the property represented by this particular value. But is this the only available interpretation of the inferred value? It may be argued that the inference referred to in Einstein's criterion only warrants the conclusion that the system in question displays a certain disposition, expressible in the following counterfactual: if we actually measured this quantity, the result would be exactly as predicted. This interpretation is more flexible than the former one, as it does not commit us to the existence of any "metaphysical" reality beyond what can be revealed in measurements. And it appears that with the counterfactual reformulation of the notion of "an element of physical reality" (or, more generally, of "property attributions") we can avoid the paradoxical consequences of the EPR argument. This result, which is unfortunately too technical to be fully presented here, comes directly from the application of the counterfactual semantics (cf. Bigaj 2006, pp. 239–246).

[4] Strictly speaking, according to the Special Theory of Relativity we should add here a phrase "with respect to an inertial frame of reference". In special relativity two events which are simultaneous with respect to some frame of reference are called "space-like separated".

Hence it may be claimed that the semantic analysis of the counterfactual can offer a tool with the help of which some conceptual difficulties of quantum mechanics can be finally laid to rest.[5]

## 3 Similarity ranking in the EPR case

We are now returning to the question posed at the beginning of the paper: which features of possible worlds should count toward their relative similarity, and which should not? Lewis acknowledges the importance of these questions, and he proposes his own celebrated solution.[6] According to Lewis, possible worlds can differ from one another with respect to two basic features: particular facts and general laws. Typically, we would consider alterations of laws of nature as producing more distant worlds than worlds constructed by alterations of individual facts only. But Lewis opposes this view. His criterion of relative similarity consists of a hierarchy of conditions, in which comparisons with respect to differences in particular facts are intertwined with comparisons with respect to law violations. Lewis presents his ranking of respects of similarity as follows:

(1) It is of the first importance to avoid big, widespread, diverse violations of law.
(2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(3) It is of the third importance to avoid even small, localized, simple violations of law.
(4) It is of little or no importance to secure approximate similarity of particular fact (Lewis 1986, pp. 47–48).

Much has been said about the feasibility of Lewis's sophisticated proposal. Some critics point out that it does not work well under the assumption of indeterminism (Percival 1999); others argue that it does not achieve the intended goal of justifying the thesis of the asymmetry of counterfactual dependence (Elga 2001). In my 2006 book I have argued that Lewis's account of similarity is prone to yet another objection. I have considered a simple example of a counterfactual statement describing a common experimental situation in quantum physics, which nevertheless cannot be correctly evaluated in the Lewis approach. The reason for this failure is not quantum indeterminism, but another non-classical feature of the quantum world, namely its non-locality. Below I shall present a sketch of my argument (for more details see Bigaj 2006, pp. 93–96).

My example involves two spin-1/2 particles (e.g. electrons) prepared in the singlet spin state for which the total spin equals 0, and two space-like separated measurements that reveal mutually opposite values of spin in a given direction for both particles. Our goal is to evaluate the counterfactual which states that if the outcome of one

[5] From what has been said, the reader can get the impression that the primary use of counterfactuals is in the context of the EPR argument. However, the same elements which admit a counterfactual reinterpretation occur in other quantum-mechanical contexts, for example in the Bell theorem (see Bigaj 2006, Chaps. 3, 6).

[6] It was presented in Lewis (1986) with corrections added as a Postscript.

measurement had been switched, the other outcome would have had to change, too. The truth of this statement seems to be an obvious consequence of the principle of the conservation of angular momentum. But it turns out that Lewis's criteria of similarity given above tell a different story. In order to evaluate our counterfactual we have to compare two possible worlds: one in which both outcomes are switched, and the other in which the distant outcome remains unchanged in spite of the local change of outcome. Although in the second world the law of the conservation of angular momentum has to be temporarily suspended, it can be argued that this suspension qualifies as a small miracle and hence enters the similarity criteria only in the third place. But an obvious advantage of admitting an instance of law violation in this case is that we can keep a large portion of space–time exactly the same as in the actual world with respect to individual facts. This is namely the infinite region that is contained within the absolute future of the distant, intact measurement, and outside the common future of both measurements (outside the intersection of two future light cones that originate at both measurements). If we look at this region in the first, law-preserving world, we can see that it may contain events that are different from that in the actual world (as they can be causal consequences of the outcome of the measurement that is different than in actuality). Hence, criterion (2) favors world two over world one, and the analyzed counterfactual turns out to be false, despite our strong intuition to the contrary.

It may be instructive to notice that one reason why Lewis's similarity criteria don't work in this case is that they are based on the prerelativistic, classical intuitions of space–time. Classically, all causal consequences of a given contrary-to-fact event $e$ affect the same area of space–time that is already affected by the occurrence of $e$, namely the entire area located "above" the absolute hypersurface of simultaneity that cuts through the location of $e$ (the classical absolute future). Consequently, eliminating a single causal consequence of $e$ with the help of a tiny miracle cannot bring any net profit in the similarity comparison, as it will not diminish the area of the (potential) discrepancy between the possible world and the actual one. On the other hand, taking into account that in the relativistic approach space-like separated points have distinct futures (forward light cones) that only partially overlap, we can see that the world in which the nomological correlation between space-like separated events is cut by a miracle will have a substantially lesser area of divergence of particular facts than the law-obeying world. In short, the failure of Lewis's method of evaluation in the EPR case is a common result of two factors: the existence of lawful correlations between space-like separated events (quantum non-locality), and the relativistic structure of space–time.

## 4 Bennett's fork theory

If we agree that the foregoing example creates a serious challenge for Lewis's analysis of the similarity relation, then the question immediately arises what to replace it with. Obviously, we are in need of a notion of counterfactual which would preserve most (if not all) of our intuitive, off-hand counterfactual judgments, while avoiding the non-locality debacle described above. One suggestion may be to abandon the notion of "miraculous events" and to work entirely within the limits of the actual laws of nature.

If we adopt this strategy, then the following approach to counterfactuals may present itself as quite natural: in order to evaluate a particular counterfactual $P\Box\rightarrow Q$, where $P$ refers to an event taking place at a certain time $T_P$, we should consider all the worlds whose temporary slice at $T_P$ contains event $P$ and all the events from the actual world that are not excluded by the existence of $P$, and which obey the usual laws of nature. In other words, we build appropriate worlds by fixing their temporal slices at $T_P$ and evolving them forward and backward in time according to the causal laws.[7] However, it turns out that this approach has many serious disadvantages.

First of all, let us notice that if we assume that the laws governing our world are strictly deterministic, the above-sketched procedure implies that worlds in which the counterfactual is to be evaluated will diverge from the actual world infinitely into the future and the past. While the difference stretching into the future does not look dangerous, the necessity of making extensive adjustments in the past certainly seems suspicious. As a consequence, we will have an infinite number of true counterfactuals whose consequents describe past events that diverge from the actual ones. Lewis refers to such counterfactuals as "backtracking", arguing energetically that the common linguistic practice allows them only in special, limited cases (Lewis 1986).

Moreover, as Bennett points out in 2003 (pp. 211–214), the Simple Theory runs the risk of pronouncing as true some counterfactuals that are clearly false according to our intuitions. Keeping all facts that occur simultaneously with the antecedent exactly as they happened in the actual world does not accord with the way we commonly use counterfactuals. Bennett correctly notices that when we utter a counterfactual supposition regarding a particular time $T_P$, we usually invoke a possible world that diverges from the actual world in a more or less "natural" way, which leads not only to the occurrence of $P$ at $T_P$, but can bring about some new facts at $T_P$ other than $P$. To use Bennett's slightly modified example: we believe in the truth of the counterfactual "If Soviet armies had reached Berlin in September 1944, it would have been heavily defended", even though actually (as far as I know) in September of the year 1944 there were few German troops in and around the capital. Notice that the presence of German defenders is not *necessitated* by the assumed fact that Stalin's advancing armies reached Berlin in September (it is logically and physically possible that the Soviets, to their enjoyment, would encounter the city virtually undefended). Hence the Simple Theory clearly forces us to reject an intuitively true counterfactual.

Incidentally, we can use the above-mentioned example as a falsifier of yet another version of Lewis's method for evaluating counterfactuals. Lewis's intricate proposal is sometimes simplified to the following form: in order to evaluate the counterfactual $P\Box\rightarrow Q$ we should consider possible worlds which are exactly as the actual world up to the time $T_P$, in which a miracle occurs just before $T_P$ that produces $P$, and which evolve from this moment on according to the usual causal laws.[8] But the example with the Soviet troops approaching Berlin in 1944 clearly shows that this method produces highly unintuitive results. For instance, we would have to accept as true that if the

[7] This method of evaluating counterfactuals has been proposed by J. Bennett in 1984. In his 2003 book Bennett rejects his previous proposal, referring to it as "the Simple Theory".

[8] Such a theory of counterfactuals is advanced by Jackson (1977).

Soviets had reached Berlin in September 1944, the Germans would still have occupied their defensive positions along the Vistula river in Poland (and hence the Soviet armies would have to have been miraculously transported over their heads to Berlin). Clearly, this consequence is not acceptable.

Bennett concludes that the agreement with our off-hand counterfactual judgments can be best achieved by postulating that the counterfactual be evaluated in possible worlds which branch off in a smooth way from the actual world so that the truth of the antecedent $P$is lawfully achieved at $T_P$. In Bennett's terminology the moment of branching is called "a fork". Bennett stresses that the fork for a given antecedent can occur much earlier than the antecedent-event. Thus, the events that would have culminated in the Soviet troops' reaching Berlin well ahead of its actual capturing time would have taken place much earlier, perhaps at the beginning of the German invasion of the Soviet Union, or maybe even earlier (one may surmise that one such fork could be Stalin's decision not to purge his best military commanders, which would lead to a dramatic increase of the Soviet army's capabilities and their stronger resistance at the beginning of the German invasion). If we place such a fork reasonably early, this would ensure that the military struggle between the two totalitarian regimes would have developed according to the standard principles of modern warfare, and thus the German defenders would have had time to fall back to Berlin as their last-ditch defense position.

Bennett generally distinguishes three possible types of forks. In a deterministic world a fork is necessarily a miraculous event, but Bennett insists that it has to be a truly insignificant miracle, not a big one (a "bump" in his terminology). As the above example shows, invoking big miracles clearly leads to the wrong valuations of commonplace counterfactuals. The second type of forks is a genuine indeterministic event, such as a quantum-mechanical coin. And, finally, Bennett maintains that sometimes in order to make sure that the counterfactual antecedent occurs at $T_P$, the differences have to stretch back in time indefinitely, but for quite a long period of time such differences may be too minute to be discernible. In such a case a fork occurs when the imperceptible differences have cumulated over time so that they are no longer negligible (Bennett calls such a situation "an exploding difference").

In should be clear that in most situations which occur in quantum mechanics we can allow ourselves the luxury of using only indeterministic forks, thus dispensing entirely with the troublesome notions of miracles and exploding differences. However, a warning is needed here. Although quantum measurement outcomes are standardly treated as indeterministic, there are interpretations of quantum mechanics (known collectively as the deterministic hidden variable theories) which assume that in fact all outcomes obtained in experiments are strictly determined by the initial, hidden state of the system. When analyzing such conceptions (even purely hypothetically, for instance when reconstructing the reasoning leading to Bell's theorem) we have to make sure that when we consider an alternative experimental result the initial state of the system is appropriately adjusted. In such a case Bennett's fork occurs not at the time of the measurement, but rather at the time of the creation of the system in question. On the other hand, a type of event that is commonly accepted as "indeterministic" in quantum mechanics is an act of selection of a given observable for measurement. Thus, if we consider a counterfactual of the form "If observable $A$ rather than $B$ had

been measured.", we can safely assume that the fork happens at the moment when the experimenter makes his decision as to what observable to measure. No adjustment of the initial state of the physical system is necessary here.

Bennett in his analysis considers an important question of whether we should put some location constraints on forks. The most natural constraint seems to be the demand that the fork leading to a given antecedent-event occur at the latest possible time. That way we would ensure that the stretch of time in which the perfect match between a given world and the actual world is maximized. However, Bennett observes that such a strict demand can produce some unwelcome consequences. If we have grounds to believe that there are two distinct ways the contrary-to-fact situation $P$ can be brought about at $T_P$, and if these ways differ in that one happens a bit earlier than the other, it would be unintuitive to arbitrarily exclude the earlier one. For instance, let us imagine that a young philosophy graduate named Frank sent two job applications—one in March to Stanford University, and one in May to Princeton University—both unfortunately without success. In this situation, without further insight into the screening procedures at both universities it is highly unlikely that the counterfactual "If Frank had received a job offer from an American university in 2006, it would have been from Princeton University" should come out true, and yet that is precisely the consequence of the strict assumption that the fork leading to a given antecedent should occur at the latest possible moment.

We can give an example showing that a similar situation can occur in quantum theory too. When three particles are prepared in the quantum state known as the GHZ state, there are three observables $X_1$, $X_2$ and $X_3$, each corresponding to one particle, and taking one of two possible values $+1$ or $-1$, such that the product of the revealed outcomes must equal $-1$ (for details see Greenberger et al. 1990; Bigaj 2006, pp. 143–145). Thus, if in an experiment the values of the observables are revealed, then if we consider counterfactually that one outcome changed its value, precisely one of the remaining two outcomes would have to change its value too. Now we can imagine a situation in which the three measurements of $X_1$, $X_2$ and $X_3$ were done in succession, and we can ask the question what would have happened had the outcome of the last measurement of $X_3$ been the opposite. Intuitively, one of the earlier measurements $X_1$ or $X_2$ would have to have its outcome changed, but if we insisted that a fork should occur at the latest possible moment, we would have to conclude that it is $X_2$, not $X_1$, whose outcome would have been switched, which seems to be unjustified. Thus, it may be reasonable to follow Bennett's advice and to pick all the worlds in which the fork occurs reasonably late, but not necessarily at the latest possible moment. Clearly, this requirement seems a bit vague, but such are our intuitions regarding the truth of some counterfactuals.[9]

[9] One theory which insists that the fork should happen at the latest possible moment is S. McCall's semantics of counterfactuals which interprets possible worlds as alternative histories (1984). In my 2006 book I similarly decided to build a semantics of quantum counterfactuals that favors later divergences from the actual world, and consequently my analysis of the GHZ example yielded the unintuitive result criticized above (see pp. 211–212). The reason for following this strategy was the formal simplicity of the resulting semantic system. Now I prefer to follow Bennett's approach to the issue of the adjustments in the past, even though it leaves certain borderline cases undecided.

## 5 The backtracking counterfactuals

The proposed method of evaluating counterfactual statements clearly implies that some backtracking counterfactuals have to be rendered true. For instance, in the above-mentioned example the counterfactual "If Frank had received a job offer in 2006, his dossier would have to have impressed the searching committee beforehand" inescapably comes out true. But how can we reconcile this fact with Lewis's strong admonishment against backtracking counterfactuals? I think that Lewis is plainly wrong when he claims that common practice almost never licenses the use of backtracking counterfactuals. The example given above proves that Lewis is undoubtedly mistaken, and similar examples are plentiful. Lewis created an impression to the contrary by a wily selection of examples in which the antecedent describes someone's action. In such cases (and, most probably, *only* in such cases) we are unwilling to accept that were the present different (because of my action), the past would have been different too, but the reason for this, as I believe, is that this counterfactual seems to presuppose that my actions are not fully free but rather conditioned by my past. And unrefined common sense clearly tells us that we are free, so no wonder that we feel reluctant to accept a statement blatantly contradicting this belief. But other cases of backtracking counterfactuals are perfectly acceptable even in ordinary discourse.

To make the case for backtracking counterfactuals even stronger, let me point out that they can be used to dismiss certain clearly paradoxical cases of counterfactual reasoning. Let us consider the following situation: suppose that a train approaches a switch at which the track diverges into two tracks which then reconverge.[10] Suppose further that the operator of the switch flips it so that the train runs down the left-hand track. This situation obviously licenses the counterfactual "Had the operator not flipped the switch, there would have been a train on the right-hand track". According to the counterfactual analysis of causation (and to our pretheoretical intuitions as well), this implies that the operator's action is a cause of the absence of the train on the right-hand track But it may be further claimed that once the train reaches the left-hand track, the following counterfactual should come out true: if there had been a train running on the right-hand track, a collision would have taken place at the point of reconvergence of both tracks. The latter counterfactual, if true, implies that the absence of a train on the right-hand track is a cause of the absence of a collision down the track. Using the assumption of the transitivity of causality we arrive at the ridiculous claim that the operator's action prevented a train collision.

I believe that the most straightforward method of dealing with this example is to deny the second counterfactual. According to the "fork" semantics of counterfactuals, in order to assess the counterfactual starting with the antecedent "If there had been a train on the right-hand track..." we have to imagine a possible world with the history identical (or almost identical) as in the actual world, and which smoothly branches off so that the antecedent is satisfied. And it should be obvious that the best way to achieve this is to assume that the operator did not move the switch, rather than assuming that a new train has been miraculously transported to the right-hand track. So by

[10] This example is borrowed from Hall (2000).

admitting a backtracking counterfactual "If there had been a train on the right-hand track, the operator would have not to have flipped the switch" we escape the apparent quagmire.[11]

## 6 Future chancy facts and similarity

So far we have agreed that the best way to approach counterfactuals within quantum theory (and within natural language too) is to consider antecedent-worlds which diverge prior to the antecedent-event (but not too early) from the actual course of history in a relatively least conspicuous way, and which subsequently evolve according to the usual laws of nature. But there is one more question that we need to address in this survey. Suppose that our actual world is partly indeterministic, as it is assumed in the standard interpretation of quantum mechanics. This means that most certainly there will be indeterministic events occurring in the future of the actual world, which may or may not be included in the closest possible antecedent-worlds. Thus the question is: should such events count toward similarity? Should we insist that the possible worlds in which to evaluate a given counterfactual be identical to the actual world with respect to all indeterministic events occurring in the latter?

Interestingly enough, there is a strong tendency to answer these questions in the affirmative.[12] This may be seen as a straightforward consequence of the fundamental feature of the Stalnaker–Lewis approach to counterfactuals, which is the postulate of the maximization of a fact-of-the-matter similarity between the relevant possible worlds and the actual world. If we can select worlds that are arguably more similar to the actual world than the others with respect to individual facts, then why not do this? However, we have to keep in mind that the ultimate test for any semantics of counterfactuals should be not its compliance with some arbitrary principles of evaluation, but rather its performance in application to real-life cases, both in natural language and in the language of any theory we are interested in (in our case this is obviously quantum mechanics). A good semantics should be able to reproduce most (if not all) truth-values associated with uncontroversial cases. Of course this does not mean that a formal analysis cannot contradict our intuitive assessments in some cases, but in such situations there should be a good explanation why this is the case.

At first sight it seems that the analysis of common-sense counterfactuals confirms that chancy future events should count toward similarity. This conclusion can be based on the evaluation of the examples of the following type. Suppose that a truly indeterministic coin has been tossed and that it has landed "heads". It is quite natural to

[11] Hall prefers a different solution to this problem, which basically denies that the truth of both forward-looking counterfactuals implies the existence of appropriate causal links. The reason for this denial is that both counterfactuals involve negative events (the absence of a train), and Hall insists that there is no such thing as causation by omission, or causation of omission. I believe that in some cases negative events are perfectly legitimate relata of the causal link, but I admit that this is a controversial issue (compare arguments against negative causation in Dowe 2004 and in favor of it in Schaffer 2004). Another possible, but no less controversial way to solve the above problem without recourse to backtracking counterfactuals is to reject the transitivity of the causal relation.

[12] See for instance Müller (2002, p. 287).

expect that the counterfactual "If I had bet on heads, I would have won" should come out true in such a situation. But this obviously implies that the world in which I placed my bet on heads and the coin fell precisely as in the actual world should be seen as more similar to the actual world than the world in which the indeterministic process of coin tossing leads to the different outcome "tails". This example seems to confirm that when evaluating a particular counterfactual we should take into account those possible antecedent-worlds whose fork occurs reasonably late, which from the moment of the occurrence of the fork evolve according to the laws of nature, *and which are identical to the actual world with respect to all chancy events that happen after the fork.*

Unfortunately, there is a high price to be paid for the achieved agreement with the indicated pretheoretical intuitions regarding counterfactuals with chancy consequents. As Bennett has proved (2003, p. 233), any counterfactual semantics which accepts that a possible world $w_1$ in which a future chancy event $E$ happens exactly as in the actual world is more similar to the actual world than a world $w_2$ in which $E$ doesn't happen implies the validity of the law of Conditional Excluded Middle (CEM):

(CEM) For all $P$, $Q$ $(P\Box\rightarrow Q)\vee(P\Box\rightarrow\sim Q)$

This consequence is unacceptable on many counts. First of all, as Lewis pointed out, there are many cases of counterfactual situations in which we have a strong inclination to reject both disjuncts in (CEM), as the famous Bizet and Verdi example illustrates. Another troublesome consequence of (CEM) is that the would- and might-counterfactuals become logically equivalent, and thus indistinguishable. And finally, with the validity of (CEM) the statement "If an observable $A$ were measured, the result would be $a$" has to be true for some value $a$. This means that counterfactuals become useless in the context of quantum mechanics, since they are not capable of expressing the fundamental feature of the quantum world which is the indetermination of the measurement outcomes. Hence it looks like we have a situation in which a lack of important theoretical virtues can outweigh the agreement with the intuitive valuations of some counterfactuals.

Bennett in his 2003 book has put forward an interesting proposal that promises to reconcile the intuitive evaluations of the counterfactuals of the form "If I had bet on heads, I would have won" with the rejection of (CEM). It turns out that in order to make sure that the considered counterfactual will come out true we don't have to unconditionally accept that all future chancy events count toward similarity. All we have to do is adopt a limited, conditional rule which prescribes, according to Bennett, that an indeterministic event counts toward similarity if it is a result of *the same causal chain* as in the actual world. Here is how this idea is supposed to work. When we evaluate the counterfactual "If I had bet on heads, I would have won" under the assumption that in the actual world an indeterministic coin tossing device produced the outcome "heads", we should consider a possible world in which numerically the same process leads to the tossing, and in which my bet was placed on "heads". Because my bet is in no way connected to the tossing, such a world will always exist (provided that we agree on some reasonable criteria of identity across possible worlds). In such a case Bennett's criterion implies that the world in which the result of the coin's tossing is identical to the actual outcome will be closer to the actual world than the world with the outcome "tails". However, Bennett's criterion is not applicable when the antecedent

of the considered counterfactual describes an event that affects the causal mechanism of the tossing device. For instance, the counterfactual "If a different person had pressed the button of the tossing device, the coin would still have landed heads" should be pronounced false, as the causal chain leading to the indeterministic outcome is plausibly a different one. And this assessment seems to go along with our intuitions nicely.

It may seem a bit odd to speak about causal chains leading to an indeterministic event, but Bennett insists that it is possible to have a cause that does not determine its effect. And I am willing to grant this to Bennett. After all, the counterfactual "If I hadn't press the button, the coin would not have landed heads" looks true (since in this case the coin would not have been tossed in the first place), so under the counterfactual interpretation of causality there is a causal link between the initiating of the device and the random outcome obtained. One may also complain about the vagueness of the notion of "identical causal chains" with respect to alternative possible worlds. Would a causal chain that started a millisecond later than in actuality count as a different one? Or a chain initiated by the button's pressing that was slightly, almost imperceptibly weaker than the actual one? To this it can be replied that with these cases we remain in the range of vagueness associated with natural language. After all, our intuitions regarding counterfactual statements are vague, as Lewis stressed several times. So it looks like Bennett's proposal is a viable alternative to the approach according to which all future chancy events should count toward similarity.

In spite of its unquestionable virtues, Bennett's causal chain approach is not my ideal solution. In my opinion it relies too uncritically on our off-hand judgments regarding a very peculiar class of counterfactuals: counterfactuals that deal with truly indeterministic processes. But we have to remember that pretheoretical linguistic intuitions can contain hidden inconsistencies, not clearly visible under a superficial analysis. It can also be the case that the intuitions we appeal to cannot be given a consistent explanation in terms of an underlying theory. I think that philosophers should not shrink from the task of giving a satisfactory explanation of the linguistic behavior of the competent users of natural language. Our linguistic intuitions should not be treated as brute facts that don't admit further explanations. And if in a particular case such an explanation cannot be formulated, this may strongly suggest that there is something wrong with the intuitive assessments regarding this case.

To return to the case in question, one can ask what is the basis of the off-hand evaluation of the counterfactual "If I had bet on heads, the result of the coin's tossing would still be heads", which implies the statement "If I had bet on heads, I would have won". I think that one possible explanation of our inclination to accept the above statement is given in the form of the underlying assumption that my wager does not affect in any way the outcome of the tossing, meaning that it cannot change the result of the tossing. But if we agree on that, we can now ask what can explain the intuitive assessment approved by Bennett, according to which the statement "If somebody else pressed the button of the tossing device, the result would still be heads" is false. It looks like an underlying assumption here is that the change of the person operating the device does influence the final outcome, so we cannot be sure that it would be the same as in the actual world. But this is plainly wrong. By assumption the device is perfectly indeterministic, so nothing we can do can have any influence on the outcome of the tossing once the random process is initiated. Neither the person who starts up

the device not the manner in which this is done can make any difference with respect to the outcome or its probability. Hence it seems that our intuitive judgment is based on the false assumption that it is somehow possible to influence the outcome of a genuinely indeterministic process.

Bennett's response to this would probably be that there is another difference between the two cases that accounts for the difference in evaluation of counterfactuals. What distinguishes the above two cases is whether the antecedent-event affects causally the process leading to the indeterministic outcome, rather than the outcome itself. But is the existence of such a causal link sufficient to account for the difference in the truth-values ascribed to the above counterfactuals? Let us suppose that the person operating the device could see me bet on a given outcome, and that this causes his hand tremble a bit. Would this seemingly insignificant fact change the intuitive assessment of the first counterfactual? I hardly think so. Even if we knew that such an effect on the tossing device can take place, we would probably be inclined toward the truth of the considered counterfactual.

However, we already know that the semantics which takes into account *all* indeterministic facts that occur in the future of the antecedent-event has very little chance of success in the context of quantum theory. This is why I propose that we ignore our unaided judgment and pronounce that *no* future indeterminate events count toward similarity. As for the common-sense examples that seem to contradict this method of evaluation, I can see several possible responses. First, we may surmise that they derive their persuasiveness from the hidden and often unnoticed presumption of determinism, or even fatalism. It can be claimed that the source of our off-hand assessment that had I bet on heads, I would have won, is the implicit supposition that the actually occurring outcome has been somehow "predestined". This belief is hidden and yet so well entrenched than even if we officially subscribe to the doctrine of indeterminism, it may still influence our judgment and lead to the fixing of the actually occurring random event.

Another way of explaining away the betting counterfactual is to respond that while its intuitive evaluation is basically correct, this fact is due to an implicit assumption that enters its antecedent along with the supposition that I bet on "heads". This is namely the assumption that the actual outcome was no other than heads. So what we are evaluating is not, strictly speaking, the counterfactual "If at time $t_1$ I had placed my bet on "heads", then at time $t_2$ I would have won", but rather " If at time $t_1$ I had placed my bet on "heads", and at time $t_2$ the coin had landed as in the actual world, then I would have won", which is an unquestionable truism.[13] On the other hand, if we wanted to evaluate the first counterfactual with proper attention paid to the temporal sequence of events, we should perform the following thought experiment: let us go back in time to the moment of a fork that leads to my betting (this fork may be safely assumed to be an indeterministic event that happens in my brain just prior

[13] It may also be speculated that the actual outcome of the indeterministic coin tossing enters the evaluation process due to the fact that the evaluation itself takes place at a later time $t_3$, hence it is tempting to fix all the occurrences prior to $t_3$ while considering the closest antecedent-worlds. However, this move is unjustified: according to the accepted semantics for counterfactuals we hold fixed the events that occurred earlier than $t_1$ (or earlier than the time of the occurrence of the fork leading to the truth of the antecedent at $t_1$).

to the betting itself) and let us evolve this situation according to the laws of nature. If we truly believe in the random character of the tossing process, we have to accept that there will be two equally matched possibilities: the coin can fall either heads or tails, so my bet is in no way guaranteed to win.

Bennett acknowledges the possibility of such a solution, but in response he creates a "scare" story which is supposed to illustrate its very unpleasant consequences (2003, p. 235). Bennett's story involves the historical event that is sometimes referred to as "the miracle of Dunkirk", namely Hitler's mysterious decision to halt the advancing German panzer divisions which gave the British troops enough respite to organize a hasty evacuation from the beaches of the French Dunkirk. Bennett proposes to assume that Hitler's decision was a purely chancy event (a result of an indeterministic firing of a couple of neurons in his brain), and under this assumption he invites us to consider the counterfactual "If one Mr. Miniver had not had a surgery at the time of the evacuation, he would have taken part in the rescue operation (as he was the owner of a large enough boat)". Bennett rightly points out that this counterfactual comes out false according to the proposed method of evaluation, as there are two equally matched possible no-surgery-worlds: one in which Hitler does what he did in the actual world, and the other in which the neurons don't fire in his brain, the order to halt the operation is not given and consequently there is no evacuation because the British troops are wiped out by the German tanks. In general, no counterfactual whose antecedent predates the evacuation and which presupposes that the evacuation took place will come out true. But our intuition seems to tell us an entirely different story—the counterfactual involving Mr. Miniver looks unquestionably true.

The above example poses a direct threat to the semantics of counterfactuals which holds that no future chancy events be taken into account when evaluating similarity to the actual world. However, I believe Bennett's theory doesn't fare much better with respect to analogous examples. True, on Bennett's interpretation the counterfactual involving Mr. Miniver comes out true, as his surgery is in no way related to the indeterministic process leading to Hitler's decision. But let us picture a different story. Suppose that in reality Hitler made up his mind regarding the halting of the offensive just after a meeting with his generals (but we are still assuming that his decision was not determined by the course of the meeting, although it was *causally connected* with the meeting). Now let us consider the counterfactual "If one of the German high command generals had been late for the meeting, the British troops would still have been evacuated from Dunkirk". I believe that our off-hand assessment of this counterfactual would be favorable as in the previous case—unless we have reasons to think that the tardiness of the general had an impact on Hitler's final decision, we tend to agree that the later events would have unfolded exactly as in the actual world regardless of our contrary-to-fact supposition. And yet the general's being late clearly affects the chain of mental events preceding Hitler's decision (for instance, Hitler could have gotten angry with the general), and hence in the closest antecedent-world the causal chain resulting in the indeterministic decision is not identical with the actual one. Consequently, according to Bennett's method the counterfactual has to be deemed false.

My opinion is that we shouldn't worry excessively about the above counterexamples. Generally speaking, our pretheoretical intuitions do not mix up well with the

assumption of indeterminism, because we have a natural inclination to look for deterministic or probabilistic causal factors even when we are told that there are none to be found. Moreover, as I have already stressed, common-sense counterfactuals are highly contextual, and this context-sensitivity is not always explicable with the help of various standards of similarity. We can imagine circumstances when it feels quite natural to reject the above counterfactuals about the Dunkirk evacuation. For instance, when various alternative scenarios of the past events are debated by historians, only those counterfactuals for which there is either a causal connection between the antecedent and the consequent, or the consequent describes an event that has already occurred at the time of the antecedent, are acceptable. But in typical situations we take advantage of the benefit of hindsight, and are willing to accept counterfactuals that don't belong to one of these two groups.

## 7 The true-antecedent counterfactuals

An interesting twist of Bennett's proposed semantics for counterfactuals is his treatment of counterfactuals with true antecedents. Those counterfactuals present a serious challenge to a semantic analysis, as there is no firm intuition associated with their use in natural language. According to Lewis's controversial solution, a counterfactual with true antecedent is true if and only if its consequent is true. Bennett opposes this move on the grounds of his distinction between deterministic and indeterministic cases. In a deterministic world Bennett agrees that when $P$ and $Q$ are true, $P\Box\rightarrow Q$ has to be true as well. However, when $Q$ is an indeterministic event, it is different. Here Bennett proposes the following criterion: if $P$ is irrelevant to the causal chain that leads to the random occurrence of $Q$, then the counterfactual is true. But if $P$ describes an event that somehow influences the causal chain leading to $Q$, the counterfactual comes out false. Hence, the counterfactual "If I had bet on heads, I would have won" uttered in the situation in which I actually bet on heads and the coin came up heads becomes true. But the counterfactual "If I had tossed the coin, it would have come up heads", whose both antecedent and consequent are true, is nevertheless false, as the antecedent is very much relevant to the process leading to the coin's landing heads.

Bennett claims that this method of dealing with true-antecedent counterfactuals is consistent with his approach to false-antecedent counterfactuals regarding future indeterministic events. And indeed it may look like this is the case. However, on closer examination we may notice that there is some sort of a tension (I hesitate to call it an inconsistency) between the intuitions underlying both treatments. Bennett's analysis of the false-antecedent counterfactual "If I had bet on heads, the coin would still have come up heads (and I would have won)" yields the truth, because the false antecedent describes an event which, if it occurred, would in no way affect the causal chain leading to the indeterministic outcome "heads". On the other hand, the true-antecedent counterfactual "If the tossing device had been turned on, the coin would have come up heads" is deemed false on the basis of the fact that the antecedent-event clearly influences the chain of events leading to the outcome "heads" (in Bennett's terminology, the antecedent-event is not irrelevant to the causal chain that leads to the outcome "heads"). But the reference to the influence exerted by the

antecedent-event in the last case is confusing. In the case of a false-antecedent counterfactual, if the antecedent-event influences the causal process leading to a chancy occurrence, this presents a threat that the outcome might be different than in the actual world, *because the antecedent-event is different than in the actual world*. But in the case of the counterfactual with true antecedent the antecedent-event is identical to that in the actual world, and its influence is presumably like in the actual world. To use Bennett's treatment of false-antecedent counterfactuals with a chancy consequent: the causal chain leading to the coin's landing heads is *identical* to the chain in the actual world, and therefore its outcome should be kept as in the actual world. Now we can definitely see a clash between the intuitions behind the two proposals Bennett put forward.

To present the case even more convincingly, we can turn once again to the false-antecedent counterfactual Bet $\Box\rightarrow$ Heads. The truth of this counterfactual is based on the hidden assumption that my betting does not influence the functioning of the coin tossing device, hence we can surmise that the counterfactual Bet $\Box\rightarrow$ Toss should be also true. But this means that in the closest Bet-world the tossing is precisely the same as in the actual world. This sameness in turn can serve as an argument for the conclusion that the outcome of the tossing should be none other than heads. But now it is extremely difficult to comprehend why the much simpler counterfactual Toss $\Box\rightarrow$ Heads should be false if the situation with respect to the tossing is precisely analogous to the situation with my bet placed on heads: we consider the closest possible world in which the same tossing process occurs, and this should lead to the same outcome "heads". Now the tension between Bennett's two methods of evaluation begins to look like a genuine inconsistency.

It looks like Bennett's semantics for false-antecedent chancy counterfactuals is more in line with Lewis's approach to true-antecedent counterfactuals. But, as I have argued against Bennett's analysis of false-antecedent counterfactuals with chancy consequents, I am not obliged to follow in Lewis's footsteps on this count. Actually, my suggested approach to counterfactuals, which involves considering our world running its course again, strongly leans towards counting all true-antecedent counterfactuals with chancy consequents as false. When evaluating both counterfactuals Toss $\Box\rightarrow$ Heads and Bet $\Box\rightarrow$ Heads, under the condition that Toss and Bet are true in the actual world (and, obviously, that Bet happens no earlier than Heads), we should consider the worlds which lead to Toss/Bet in the least conspicuous way, and which from that moment on follow the usual laws of nature. This means that the evaluation of both counterfactuals is done in the worlds which look exactly like our world up to the time of Toss, and which then can choose any path that agrees with the causal laws. And because we have assumed that the outcome of the coin tossing is truly indeterministic, we are left with both counterfactuals rendered false.

I believe that the suggested method of evaluating counterfactuals nicely deals with typical cases that can be encountered in quantum mechanics. To illustrate this, let us suppose that a quantum system $s$ is prepared in state $\varphi$ that is a non-trivial superposition of eigenstates for a particular observable $A$. In such a case the counterfactual "If I measured $A$, the outcome would be $a$" is false for any value $a$, which reflects the objective indeterminacy of the physical property $A$. Now let us assume that actually at a moment $t$ the measurement of $A$ was performed, and that the revealed outcome happened to be $a$. According to the standard interpretation of measurement, the initial

state of the system $\varphi$ reduces upon measurement to the eigenstate $\varphi_a$ corresponding to the value $a$. The counterfactual "If I measured $A$ again after $t$, the outcome would be $a$" now becomes true, as we have to keep the past of the second measurement (which is assumed to be an indeterministic event) exactly as in the actual world, and in the actual world the first measurement revealed $a$. On the other hand, our evaluation looks different if we consider the true-antecedent counterfactual "If I measured $A$ at $t$ (which I actually did), the outcome would be $a$ (which it was)". Here we have to imagine a second run of the measurement in the case the initial state is $\varphi$, not $\varphi_a$, and therefore all options are still on the table: the outcome may, but doesn't have to, be $a$ again. Hence the counterfactual is false, and this reflects the fact that at the beginning of the actual measurement the system is not yet determined with respect to the quantum property $A$.

Finally, we can consider the following false-antecedent counterfactual "If I sneezed at $t$, the outcome would be $a$", under the condition that in the actual world the measurement revealed $a$. Our inclination to the contrary notwithstanding, I strongly suggest that this counterfactual be treated as false. We have to resist the temptation to interpret this fact as an indication that there is a mysterious interaction between my sneezing and the measuring device, which affects the revealed outcome. Rather, the interaction would be present if we had a counterfactual dependence here, i.e. if the counterfactual "If I sneezed, the outcome would not be $a$" were true. But it is obviously false, and this indicates only that in a new counterfactual situation no outcome is yet determined as of $t$.

To sum up, I argue that indeterministic facts which occur after the antecedent-event are entirely irrelevant to the evaluation of quantum counterfactuals. In order to asses the truth-value of a given counterfactual, we have to select all possible worlds which start off like the actual world does, in which a reasonably late fork leads to the truth of the antecedent, and whose evolution from that moment on is governed by the actual causal laws. One aspect of this evaluation method that is still left untouched is its postulated relativistic invariance. It turns out that the separation of space–time into the part "before" the fork which is supposed to be the same as in the actual world, and the part "after" the fork, the only constraints on which is its compliance with the laws of nature given the initial conditions, is not uniquely determined by the principles of the special theory of relativity. In my 2004 and 2006 I present in details two possible ways of drawing this separation, using the forward and backward light cones (see also Finkelstein 1999). But regardless of their differences, both semantics of quantum counterfactuals follow the general principles that have been argued for in this article.



## References

Bedford, D., & Stapp, H. P. (1995). Bell's theorem in an indeterministic universe. *Synthese, 102*, 139–164.
Bennett, J. (1984). Counterfactuals and temporal direction. *The Philosophical Review, XCIII*(1), 57–91.
Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Clarendon Press.

Bigaj, T. (2004). Counterfactuals and spatiotemporal events. *Synthese, 142*(1), 1–20.
Bigaj, T. (2006). *Non-locality and possible worlds: A counterfactual perspective on quantum entanglement*. Frankfurt: Ontos Verlag.
Bigaj, T. (2007). Counterfactuals and non-locality of quantum mechanics. *Foundations of Science, 12*, 85–108.
Clifton, R., Redhead, M., & Butterfield, J. (1990). Nonlocal influences and possible worlds. *British Journal for the Philosophy of Science, 41*, 5–58.
Dowe, P. (2004). Causes are physically connected to their effects: Why preventers and omissions are not causes. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 189–196). Oxford: Blackwell.
Einstein, A., Podolsky, B., & Rosen, N. (1935).Can quantum-mechanical description of physical reality be considered complete? *Physical Review, 48*, 696–702.
Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science, 68*, S313–S324.
Finkelstein, J. (1999). Space–time counterfactuals. *Synthese, 119*, 287–298.
Ghirardi, G., & Grassi, R. (1994). Outcome predictions and property attribution: The EPR argument reconsidered. *Studies in the History and Philosophy of Science, 25*(3), 397–423.
Greenberger, D., Horne, M., Shimony, A., & Zeilinger, A. (1990). Bell's theorem without inequalities. *American Journal of Physics, 58*, 1131–1143.
Griffiths, R. B. (1999). Consistent quantum counterfactuals. *Physical Review A, 60*(1), R5–R8.
Griffiths, R. B. (2001). *Consistent quantum theory*. Cambridge: Cambridge University Press.
Hall, N. (2000). Causation and the price of transitivity. *The Journal of Philosophy, 97*(4), 198–222.
Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy, 55*, 3–21.
Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.
Lewis, D. (Ed.) (1986). Counterfactual dependence and time's arrow. In *Philosophical papers* (Vol. II). Oxford: Oxford University Press.
McCall, S. (1984). Counterfactuals based on real possible worlds. *Nous 18*, 463–477
Mermin, D. (1998). Nonlocal character of quantum theory?. *American Journal of Physics, 66*(10), 920–924.
Müller, T. (2002). Branching space-time, modal logic and the counterfactual conditional. In T. Placek & J. Butterfield (Eds.), *Non-locality and modality* (pp. 273–291). Dordrecht: Kluwer.
Percival, P. (1999). A note on lewis on counterfactual dependence in a chancy world. *Analysis, 59*(3), 165–173.
Redhead, M. (1987). *Non-locality, incompleteness and realism*. Oxford: Oxford University Press.
Schaffer, J. (2004). Causes need not be physically connected to their effects: The case for negative causation. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 197–216). Oxford: Blackwell.
Shimony, A., & Stein, H. (2001). Comment on 'nonlocal character of quantum theory' by Henry P Stapp. *American Journal of Physics, 69*, 848–853.
Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*. Oxford: Blackwell.
Stapp, H. P. (1971). S-matrix Interpretation of quantum theory. *Physical Review, 3*, 1303–1320.
Stapp, H. P. (1989). Quantum nonlocality and the description of nature. In J. Cushing & E. McMullin (Eds.), *Philosophical consequences of quantum theory reflections on Bell's theorem* (pp. 154–174). Notre Dame: University of Notre Dame Press.
Stapp, H. P. (1997). Nonlocal character of quantum theory. *American Journal of Physics, 65*, 300–304.
Stapp, H. (1998). Meaning of Counterfactual Statements in Quantum Physics. *American Journal of Physics, 66*(10), 924–926.
Stapp, H. P. (2001). Bell's theorem without hidden variables. Lawrence Berkeley Laboratory Report No. LBNL 46942, quant-ph/0010047.
Vaidmann, L. (1999). Defending time-symmetrized quantum counterfactuals. *Studies in History and Philosophy of Modern Physics, 30*, 373–397.