Tomasz Bigaj

Counterfactuals and non-locality of quantum mechanics

Abstract

In the paper the proof of the non-locality of quantum mechanics, given recently by H. Stapp and D. Bedford, and appealing to the GHZ example, is analyzed. The proof does not contain any explicit assumption of realism, but instead it uses formal methods and techniques of the Lewis calculus of counterfactuals. To ascertain the validity of the proof, a formal semantic model for counterfactuals is constructed. With the help of this model it can be shown that the proof is faulty, because it appeals to the unwarranted principle of "Elimination of Eliminated Conditions" (EEC). As an additional way of showing unreasonableness of the EEC assumption, it is argued that yet another alleged and highly controversial Stapp's proof of non-locality of QM, using the Hardy example, can be made almost trivial with the help of EEC. Next the question is considered whether the validity of the proof in the GHZ case can be restored by adopting the assumption of "partial realism". It is argued that although this assumption makes the crucial step of the original reasoning valid, it nevertheless renders another important transition unjustified, therefore the entire reasoning collapses.

Introduction

One of the greatest achievements of the foundational analysis of quantum mechanics is undoubtedly the Bell theorem. Although there is still a debate going on concerning particular consequences of this theorem, it is clear that its main message is the following: there can be no hidden-variable theory which would produce the same observational consequences as standard quantum mechanics, and yet would be local. The structure of the original Bell argument is well known: it proceeds from the assumption of the existence of a hidden variable theory (or in other words from the ontological assumption of realism about values of observables) and from the suitably formulated assumption of locality (roughly speaking, this assumption says that a distant measurement performed on one system cannot change the objective value of an observable pertaining to another system) to a consequence which is incompatible with the standard quantum-mechanical predictions (this consequence is the Bell inequality). Since the time of the original Bell argument there have been numerous attempts to generalize or to improve it with respect to both technical details and fundamental assumptions. Hence now we have Bell-like theorems which deal with imperfect correlations or imperfect measurements, which refer to more than two separate particles, which don't involve any inequalities, etc.[1]

Without doubts, philosophically the most interesting attempts are those that aim at weakening the crucial assumptions of the original Bell theorem. A notable example of this kind of undertakings can be found in works of Henry P. Stapp, a physicist with a strong philosophical motivation. Since 1971 he has published several papers in which he argues essentially that the assumption of realism is not necessary in obtaining a contradiction with quantum-mechanical predictions. Stapp claims that from the locality assumption alone together with quantum-mechanical predictions a contradiction can be derived. This claim, if justified, would constitute the strongest argument for the non-locality of quantum mechanics.[2] What is particularly interesting about Stapp's arguments is that he relies heavily on counterfactual reasonings, i.e. reasonings concerning what would have happen, had things been different. This fact of course can raise justified suspicions. After all, aren't we convinced by the founding fathers of quantum mechanics (like Bohr himself) that within quantum realms we should talk only about what *is* measured, and not what *could*, or *would* be measured? This suspicion toward counterfactuals is explicitly expressed for example by B. van Fraassen (1991), when he says that

> the violation of Bell's inequalities demonstrates empirically that we should not look to measurement outcomes to give us direct information about state, propensity, capacity, ability, or *counterfactual facts*. From fact to modality, only the most meager inferences are allowed (p. 125, italics mine).

Yet there is a strong opposition against this "positivistic" approach to the validity of counterfactual suppositions in quantum theory. Many physicists and philosophers point out that whenever we talk about laws of nature (and quantum mechanics does not drop this talk at all), we must assume some kind of "counterfactuality": laws tell us not only what happens actually, but first and foremost, what would happen, if the conditions were such-and-such (see Maudlin 1994, pp. 126-32; Unruh 1999). The only problem with counterfactuals is that we should be extremely cautious in using them. For example, as van Fraassen rightly points out,

---

[1] For an extensive but accessible survey of different versions of Bell's theorem, see (Placek 2000, chapter 5 "Charting the labyrinth of the Bell's theorems").

[2] It should be stressed here that by "non-locality" Stapp means something really strong: not only that there are correlations between results of experiments on spatially separated particles (which is well known and widely accepted), but that the mere performing of measurement on one system can influence the result of the experiment on a distant system (these two types of non-locality were originally introduced in Jarrett 1984; after A. Shimony it is common to call them "outcome dependence" and "parameter dependence").

we cannot take for granted some apparently plausible rules, like the rule of counterfactual definiteness, or "counterfactual excluded middle", which says that for any sentences *A* and *B*, either if it were that *A*, then it would be *B*, or if it were that *A*, then there would be not *B*.

It is commonly held that the Lewis formal analysis of counterfactuals, inspired by Stalnaker's approach but differing from it in some crucial points, can overcome these difficulties and be safely applied to quantum-mechanical phenomena. This is the assumption which is made by Stapp. He claims to have used the Lewis counterfactuals in order to prove his theorem. Actually, three different versions of his argument can be found in his writings. The earliest version, using the original EPR-like situation from the Bell theorem, was presented in (Stapp 1971), and then refined on different occasions (Stapp 1989). This argument was heavily criticized by Redhead and the others (Redhead 1987, Redhead *et al*. 1990). Another, most recent argument, uses an example known as the Hardy experiment (Stapp 1997). It ignited a lively discussion involving philosophically-oriented physicists (Mermin 1998, Unruh 1999, Finkelstein 1998). And there is a third version of Stapp's theorem, dealing with the so-called GHZ example. This argument was originally formulated in (Stapp 1991) and then refined and formalized in (Bedford, Stapp 1995) with explicit reference to the Lewis calculus of counterfactuals.

The GHZ version of Stapp's argument, especially in its version from 1995, is definitely the most advanced with respect to logical details. Maybe for that reason it has not received much attention from physicists and philosophers, and has passed virtually unchallenged. Yet it appears that in spite of its meticulous presentation with respect to almost every technical detail, Stapp's proof[3] contains serious flaws which make the conclusion unjustified. In what follows, I would like to present a detailed analysis of this proof, which can serve different purposes. First of all it will clarify somewhat obscure steps of the argument, by pointing out that this argument is a counterfactual version of a much simpler argument which can be made when the assumption of realism is invoked, much as in the original Bell theorem. But most importantly, my analysis is going to show where exactly Stapp's argument fails, and why his conclusion is unwarranted. I will also raise the question whether the argument can be corrected by adding an additional assumption which I call "partial realism". It will turn out that in spite of its initial plausibility, even this partial realism cannot produce the needed results. As far as the technical aspect of my article is concerned, I am going to introduce certain semantic tools for assessing validity of logical transitions within

[3] With all required credits for the second co-author, D. Bedford, I will call the analyzed argument for short "Stapp's argument" throughout the paper.

Stapp's alleged proof. These tools will be essentially Lewis-style possible-world semantic models for counterfactuals, tailored to the physical situation in question.

1. Initial assumptions

The original proof deals with an example of the Greenberger–Horne–Zeilinger (GHZ) situation in its version with three particles. I will not go here into the physical details concerning this example (for a physical presentation see Greenberger *et al.* 1990; Auletta 2000, pp. 628-33). The only thing we should know is that there are three particles 1, 2, 3, and each of them can undergo one of two measurements X_i or Y_i, with two possible results each: x_i = +1 or –1 and y_i = +1 or –1. When these three particles are prepared in a special initial state, there are certain correlations between possible results of the above-mentioned measurements. These correlations can be presented as follows:

(QM1) $X_1X_2X_3 \Rightarrow x_1x_2x_3 = -1$

(QM2) $X_1Y_2Y_3 \Rightarrow x_1y_2y_3 = +1$

(QM3) $Y_1X_2Y_3 \Rightarrow y_1x_2y_3 = +1$

(QM4) $Y_1Y_2X_3 \Rightarrow y_1y_2x_3 = +1$

In words (QM1–4) says that when the measurements X_1, X_2, and X_3 are jointly performed, the product of their results must equal –1, and with the other three combinations the product is positive. It is important to stress that these correlations take place even if measurements are space-like separated from each other, so that no known physical interaction can appear between all particles.

Now, it is quite obvious that the above predictions lead to a contradiction when we assume that each observable in question has its objective value independently of the measurement revealing it, and independently of the two other measurements – in other words, that all numbers x_i and y_i have determinate values. To prove this, it suffices to multiply all sides of the equations (QM1-4). On the right-hand side we will obtain the product –1, but on the left-hand side each particular value will be squared, and therefore their total product must equal 1.

Here we have repeated the usual Bell result that realism plus locality is incompatible with standard formalism of QM. However, Stapp claims that with the help of this GHZ example we can prove even more: without assuming realism, only using counterfactual reasonings about possible experiments and assuming some version of locality, he wants to show that a contradiction can be derived from (QM1–4). Actually, his proof of this contradiction, both in the earlier and in the refined version, consists of a quite complicated and not necessarily intuitive chain of derivations. But we can make it a little bit more accessible by pointing out that this proof essentially mimics one of several possible ways of deriving a contradiction in the easier case of realism. However, this is not the way we have just presented. Obviously, the reasoning sketched above requires multiplication of six different results of experiments, and we cannot hope to reproduce this in counterfactual reasoning. We should rather find such a way of proceeding that at each step only a minimal number of different values is invoked. And here is one such possible way of reasoning.

We start, as before, with the assumptions (QM1) and (QM2), noting the following implication:

(R1) $x_2x_3 = -p \Rightarrow y_2y_3 = p$

This implication goes through, because we assume the existence of the objective value of x_1. In the same way we can proceed from (QM3) and (QM4):

(R2) $y_2x_3 = q \Rightarrow x_2y_3 = q$

And now suppose that three observables X_2, X_3 and Y_2 have the following values: $x_2 = m$, $x_3 = n$, and $y_2 = r$. Then by (R1) we have $y_2y_3 = -mn$, and hence

(R3) $x_2 = m \wedge x_3 = n \wedge y_2 = r \Rightarrow y_3 = -mnr$

But from (R2), which is true for all q, we can obviously infer that $x_2y_3 = rn$, which means, given (R3), that $x_2 = -m$. Here we have ended up with a contradiction: from the assumption that $x_2 = m$ we have derived that $x_2 = -m$.

Now we can at least hope to find counterfactual representations for each step in this reasoning. For example step (R1) could be presented as follows: if we measured X_1, X_2, X_3, and obtained $x_2x_3 = p$, then if we had chosen Y_2Y_3 instead of X_2X_3, we would have obtained $y_2y_3 = -p$. However, the proof of this counterfactual counterpart will turn out to be somewhat intricate, to put it mildly.

Stapp obviously must rely on some assumptions. His main premise is the assumption of locality, interpreted with the help of counterfactual conditionals. It says roughly that when we change counterfactually one or two of the measured observables, the result obtained in the third measurement should not change. We will take this assumption for granted, although it was heavily criticized in the context of another of Stapp's arguments by Redhead[4]. Then there is an entire battery of valid patterns of inference in the Lewis counterfactual calculus, whose validity Bedford and Stapp establish scrupulously in (1995). And the third component of Stapp's auxiliary premises consists of two patterns of inference, which although not generally valid in the Lewis calculus, are claimed to be valid in the particular context of the GHZ example. The first rule is called "Elimination of Eliminated Conditions" (note the tautological character of this nomenclature); the second has no name, but it can be called "Addition of Irrelevant Conditions". I will formulate them later, when they are needed. Not surprisingly, it will appear that one of them is essentially responsible for the failure of the entire reasoning.

The first step in the proof aims at showing the validity of the counterfactual analogue of the thesis (R1):

(C1) $X_2X_3 \wedge x_2x_3 = -p \Rightarrow (Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = p)$

It appears that this is not such an easy task. First, we will have, following Stapp, to appeal to the locality assumption:

(LOC1) $X_1X_2X_3 \wedge x_1 = p \Rightarrow (X_1Y_2Y_3 \;\Box\!\!\rightarrow x_1 = p)$

In words: if we obtain the result $x_1 = p$, while choosing for the other two particles measurements X_2 and X_3, then this result should be valid even if we counterfactually chose Y_2 and Y_3. Now, when we appeal to the predictions (QM1) and (QM2), we can easily convince ourselves that the following must hold:

(1) $X_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (X_1Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = p)$

[4] The essential point of disagreement between Stapp and Redhead lies in a different way of reading counterfactuals with antecedents referring to a localized spatiotemporal event. Stapp implicitly assumes that in order to evaluate such a counterfactual we should analyze all possible worlds which are the same as the actual one with respect to the whole space-time region outside the absolute future of the event, whereas Redhead claims that what should be kept fixed is confined only to the absolute past of this event. I have analyzed this distinction extensively elsewhere (Bigaj 2002a, 2002b).

But (1) still falls short of the needed (C1). In (1) the counterfactual argument from $x_2x_3 = –p$ to y_2y_3 goes through only in virtue of the measurement X_1 being an "intermediate" element. But we need something stronger: no matter what measurement is performed on the particle 1, as long as $x_2x_3 = –p$, the results of the would-be measurements Y_2 and Y_3 must obey the equation $y_2y_3 = p$. And here Stapp tries the following route: suppose that the actual measurement performed on the particle 1 is Y_1. In virtue of the locality assumption we can argue for the following:

(LOC1′) $Y_1X_2X_3 \wedge x_2x_3 = –p \Rightarrow (X_1X_2X_3 \;\Box\!\!\rightarrow x_2x_3 = –p)$

Now we can proceed using the already proven (1), and replacing the consequent of the counterfactual by the consequent of (1) (this move is in agreement with the Lewis rules of inference):

(2) $Y_1X_2X_3 \wedge x_2x_3 = –p \Rightarrow (X_1X_2X_3 \;\Box\!\!\rightarrow\; (X_1Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = p))$.

And in order to return to the situation when $Y_1Y_2Y_3$ are performed, we can still appeal to the locality condition, arguing that the equation $y_2y_3 = p$ should remain unchanged. Hence we obtain the following chain of counterfactuals:

(3) $Y_1X_2X_3 \wedge x_2x_3 = –p \Rightarrow (X_1X_2X_3 \;\Box\!\!\rightarrow\; (X_1Y_2Y_3 \;\Box\!\!\rightarrow (Y_1Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = \mathrm{p})))$.

We can see that the first and the last elements of this chain are exactly the ones we need to get, if we want to obtain a version of (1) with the measurement Y_1 instead of X_1. But how to get rid of the intermediate elements (intermediate counterfactual situations)? Here Stapp appeals to the earlier announced principle of the "Elimination of Eliminated Conditions". Essentially, he would claim that each counterfactual supposition in (3) annuls the preceding one, so we are finally left with the last only. This may look convincing at first sight, but let us look closer. First consider the implication (2) and ask if we are allowed to cross out from it the counterfactual condition $X_1X_2X_3$. This might seem difficult to answer, because "double counterfactuals" are quite non-intuitive. But we can restate (2) in terms of possible worlds, keeping in mind the ordinary truth-conditions for counterfactuals as imposed by Lewis. Let w_0 be the "actual" world, i.e. the world in which $Y_1X_2X_3$ are performed, and the product x_2x_3 equals $-p$. Let w_1 represent the world closest to the actual, in which $X_1X_2X_3$ are performed. This is the world in which x_2x_3 still equals $-p$, and therefore by (QM1) $x_1 = p$. Now, in order for (2) to be true, the counterfactual $X_1Y_2Y_3 \Box\!\!\rightarrow y_2y_3 = p$ must be true at w_1. This, on the other hand, means that when we take the world w_2 closest to w_1 and such that $X_1Y_2Y_3$ holds, then in this world the consequent $y_2y_3 = p$ should hold. Because w_2 is closest to w_1, x_1 should be the same as at w_1, therefore at w_2 $x_1 = p$ and by (QM2) we have $y_2y_3 = p$. In that way we can argue that (2) is indeed true, if we assume the ordinary truth-conditions for counterfactuals, together with some reasonable rules for comparing similarity between possible worlds.

2. Semantic models for counterfactuals

Because things are getting here quite complicated, and because we will have to assess validity of certain "non-intuitive" transformations, it would be a good idea to proceed slightly more formally. Stapp's entire argument relies on the standard Lewis truth-conditions of counterfactuals and on some rules of comparative similarity between possible worlds (see Lewis 1973, 1986). Therefore we can now formally construct a semantic model, consisting of a set of possible worlds and of some rules defining the relation of comparative similarity between these worlds. These rules will be incomplete, for reason which will become clear soon, but sufficient for the valuations of all of Stapp's transformations. Let us first start with the definition of the set of possible worlds. In our case of the GHZ example a given possible world is fully defined by specifying three measurements and their results. Therefore we will formally represent a possible world by a sextuple $\langle Z_1, Z_2, Z_3, z_1, z_2, z_3\rangle$, where $Z_i = X_i$ or Y_i and $z_i = +1$ or -1. In general, the experimental situation allows for $2^6 = 64$ different possible worlds, but because of the restrictions QM1–4 we have in fact 16 worlds less, therefore the final number of possible worlds is 48.

Now we have to introduce some rules of comparative similarity. Let $w_0 = \langle Z_1^0, Z_2^0, Z_3^0, z_1^0, z_2^0, z_3^0\rangle$ be the actual world, and let $w_1 = \langle Z_1^1, Z_2^1, Z_3^1, z_1^1, z_2^1, z_3^1\rangle$ and $w_2 = \langle Z_1^2, Z_2^2, Z_3^2, z_1^2, z_2^2, z_3^2\rangle$ be two possible worlds. In comparing w_1 and w_2 with respect to their closeness to w_0, we should take into account both the measurements performed and the results obtained. Let us define $\mathcal{Z}^{10} = \{Z_i^1 : Z_i^1 = Z_i^0\}$, i.e. $\mathcal{Z}^{10}$ is a set of measurements performed in the world w_1, which are the same as in the actual one. Analogously, $\mathcal{Z}^{20} = \{Z_i^2 : Z_i^2 = Z_i^0\}$. The first partial rule of comparative similarity, suggested by Stapp's remarks, will be the following:

(CS1) If the number of elements in $\mathcal{Z}^{10}$ is no less than in $\mathcal{Z}^{20}$, then if the number of measurements in $\mathcal{Z}^{10}$ with the same result as in w_0 is greater than the number of measurements in $\mathcal{Z}^{20}$ with the same results as in w_0, then $w_1 <_0 w_2$.

The expression "$w_1 <_0 w_2$" is an abbreviation for "w_1 is closer to w_0 than w_2". The rule (CS1) says that the number of the measurements with the same results as in the actual world counts towards similarity, provided that the number of repeated measurements is not decreased. (CS1) already implies Stapp's version of the locality assumption, because according to it we should always judge as closer to reality the world in which the result of an unchanged measurement is the same, even if the other measurements had been chosen different.

The second rule shows that in some cases the mere difference in numbers of the same measurements as in the actual world can count towards similarity.

(CS2) If the number of measurements in $\mathcal{Z}^{10}$ with the same result as in w_0 is no less than the number of measurements in $\mathcal{Z}^{20}$ with the same result as in w_0, then if the number of elements in $\mathcal{Z}^{10}$ is greater than the number of elements in $\mathcal{Z}^{20}$, then $w_1 <_0 w_2$.

(CS2) implies another version of the locality assumption; namely that when we change some measurement settings, the remaining settings should be unchanged. We can also add the third rule of comparative similarity to the effect that the only situation in which we are allowed to assert $w_1 <_0 w_2$ is when the number of repeated measurements or the number of repeated measurements with the same results is greater in w_1 than in w_2.

(CS3) If $w_1 <_0 w_2$, then either the number of elements in $\mathcal{Z}^{10}$ is greater than the number of elements in $\mathcal{Z}^{20}$, or the number of measurements in $\mathcal{Z}^{10}$ with the same result as in w_0 is greater than the number of measurements in $\mathcal{Z}^{20}$ with the same results as in w_0.

A consequence of (CS3) is that the results of counterfactually altered measurements do not by themselves count towards similarity, which seems reasonable.

It should be quite obvious that rules (CS1-3) are not sufficient to determine in each case whether one world is closer to the actual than the other. Namely, the rules presented above don't decide what is more important for comparative similarity: the number of repeated results, or the number of repeated measurements. For example, when the actual world is the following: $w_0 = \langle X_1, X_2, Y_3, -1, -1, -1\rangle$, then our rules of comparative similarity cannot help in assessing which world is closer to w_0: $w_1 = \langle X_1, X_2, X_3, +1, +1, -1\rangle$, or $w_2 = \langle Y_1, Y_2, Y_3, +1, -1, -1\rangle$. In w_1 two of three measurements are the same as in w_0, but none of them with the same result; in w_2 the number of repeated measurements is lesser, namely one, but the result of this repeated measurement is the same as in w_0. But (CS1) and (CS2) are completely sufficient to evaluate the entire reasoning presented by Stapp.[5]

Because our universe consists of finitely many possible worlds, we can use the following, simpler version of the Lewis truth conditions for counterfactuals:

(TC) The counterfactual $A \,\square\!\!\rightarrow B$ is true at the world w_0 iff B is true at all A-worlds closest to w_0.

Let us apply this semantic model to asses again the validity of the statement (2). The strict implication is true if its consequent is true at all possible worlds fulfilling the antecedent. In other worlds: the counterfactual $X_1X_2X_3 \,\square\!\!\rightarrow (X_1Y_2Y_3 \,\square\!\!\rightarrow y_2y_3 = p)$ must be true at all following worlds: $\langle Y_1, X_2, X_3, \_, p, -p\rangle$, $\langle Y_1, X_2, X_3, \_, -p, p\rangle$, where "_" stands for any possible result. That means, according to the rule (CS1), that in every world of the type $\langle X_1, X_2, X_3, \_, p, -p\rangle$ or $\langle X_1, X_2, X_3, \_, -p, p\rangle$, the sentence $(X_1Y_2Y_3 \,\square\!\!\rightarrow y_2y_3 = p)$ must hold. But because of the quantum-mechanical prediction QM1, the blank in the result of the X_1-measurement should be replaced by p. Therefore, again according to (CS1), the closest $X_1Y_2Y_3$-worlds to these worlds are the following: $\langle X_1, Y_2, Y_3, p, \_, \_\rangle$. But again using the quantum-mechanical prediction QM2 we see that the product of the two blanks must equal p, and therefore the validity of (2) is proven.

[5] One possible way of dealing with this situation is to admit that some possible worlds are not comparable at all. As a consequence, the relation of comparative similarity would be no longer a linear ordering, but only a partial ordering. To see that this might be the case, observe that we cannot for example pronounce the worlds w_1 and w_2 as equisimilar, and by this assure fulfillment of the requirement of the connectivity. For in that case we could introduce a world $w_3 = \langle X_1, X_2, Y_3, +1, +1, +1\rangle$, which by the same assumption should be equally similar to w_2, and yet according to (CS2) $w_3 > w_1$, which leads to a contradiction. A similar argument in favor of the thesis that in the relativistic context comparative similarity must be a partial ordering is formulated in (Finkelstein 1999). For the consequences of this thesis for the truth-conditions of counterfactuals see (Bigaj 2002b).

But what about the validity of (2) with $X_1X_2X_3$ crossed out? Let us write it down:

(4) $Y_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (X_1Y_2Y_3 \,\Box\!\!\rightarrow y_2y_3 = p)$

Now the truth-conditions imply the following: if (4) were to be true, $y_2y_3 = p$ must hold at all the worlds $\langle X_1, Y_2, Y_3, \_, \_, \_\rangle$ which are closest to some of the worlds $\langle Y_1, X_2, X_3, \_, -p, p\rangle$ or $\langle Y_1, X_2, X_3, \_, p, -p\rangle$. But because no measurement in the former is the same as in the latter, according to (CS3) we can impose no condition on what the results of the measurements $X_1Y_2Y_3$ should be in $\langle X_1, Y_2, Y_3, \_, \_, \_\rangle$ that are closest to some "actual" world. Therefore there will be one such a world in which y_2y_3 will not be equal to p, and hence counterfactual (4) will come out false. The transition from (2) to (4) is definitely not valid.

The rule of the "Elimination of Eliminated Conditions" cannot be taken for granted. Stapp formulates it in the following way: if M_1, M_2 and M_3 are three alternative triplets of measurements, φ is a possible outcome of M_1 and $P(\varphi)$ is a proposition that "depends only on φ", then the following pattern of inference is valid:

(EEC) If $M_1 \Rightarrow (M_2 \,\Box\!\!\rightarrow (M_3 \,\Box\!\!\rightarrow P(\varphi)))$, then $M_1 \Rightarrow (M_3 \,\Box\!\!\rightarrow P(\varphi))$

I don't know what the phrase "depends only on φ" exactly means. It could mean that P refers only to the result of the measurement of M_1, but this would be unreasonable, because the antecedent of the last counterfactual assumes M_3, not M_1. And besides, in the application leading from (2) to (4), the sentence $y_2y_3 = p$ definitely does not refer to the result of the actual measurement. For that reason I will ignore this restriction on P. And now we can see that (EEC) is unwarranted. The counterfactual assumption M_2 may introduce some new elements to the whole situation, which together with another counterfactual assumption M_3 can lead to P, but nonetheless P will not follow from the counterfactual assumption M_2 alone. It seems to me that Stapp overlooked the fact that the "annulment" of the assumption M_2 by M_3 is not total – we assume that some of our choices are different, but some remain the same, and in virtue of the meaning of counterfactuals we have to leave intact as many facts as possible from the world in which M_3 holds.

In the conditional (2) the element which guarantees the truth of the consequent is of course the result of the measurement X_1. And now it can be conjectured that in claiming the validity of the move from (2) to (4), Stapp implicitly assumes the objective reality of the value

x_1, in spite of the fact that no measurement was performed to reveal this value. This observation can serve as a possible explanation of why Stapp was able to derive a contradiction: he apparently included some residual form of reality assumption, at least with respect to the non-measured observable X_1. But it will appear later that even under this assumption of "partial realism" Stapp's final conclusion is unwarranted.

3. A Hardy-type experiment

In order to see how powerful and therefore unwarranted the assumption (EEC) is, let us observe that with the help of it another proof of non-locality by Stapp (1997) can be made almost trivial. This proof concerns a Hardy-type experiment with two particles *L* and *R*, mentioned earlier. I will present here the self-explanatory diagram, together with some clarifying remarks.

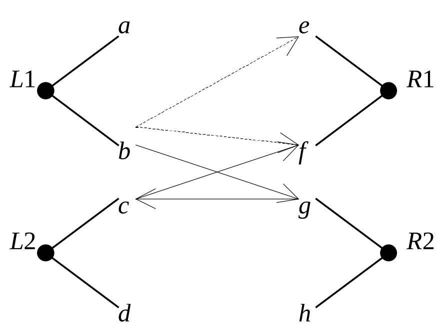

Fig. 1. Quantum-mechanical predictions concerning the Hardy experiment. Solid arrows indicate strict implications, dashed arrows indicate "possibility". For example, the arrow leading from the result *b* of the experiment *L*1 to the result *g* of the experiment *R*2 shows that when *L*1 is performed and the result obtained is *b*, the result of *R*2 must be *g*. Dashed arrows leading from *b* to *e* and *f* indicate that when *L*1 gives *b*, both results of the measurement *R*1 have non-zero probability.

And now I argue that the following chain of counterfactual reasonings must be true:

(H1) $L1 \wedge b \wedge R2 \Rightarrow (L2 \wedge R2 \;\Box\!\!\rightarrow L2 \wedge R2 \wedge c)$ (by QM predictions and Stapp-locality LOC1)

(H2) $L2 \wedge c \wedge R2 \Rightarrow (L2 \wedge R1 \;\Box\!\!\rightarrow L2 \wedge R1 \wedge f)$ (by the same as above)

(H3) $L2 \wedge R1 \wedge f \Rightarrow (L1 \wedge R1 \;\Box\!\!\rightarrow f)$ (by the locality alone)

Combining these three together we obtain

(H4) $L1 \wedge b \wedge R2 \Rightarrow (L2 \wedge R2 \;\Box\!\!\rightarrow (L2 \wedge R1 \;\Box\!\!\rightarrow (L1 \wedge R1 \;\Box\!\!\rightarrow f)))$

By appealing to (EEC) we could get rid of the intermediate elements, thus being left with

(H5) $L1 \wedge b \wedge R2 \Rightarrow (L1 \wedge R1 \;\Box\!\!\rightarrow f)$

This is exactly step (6) in the original Stapp proof. From this we can obtain a contradiction with predictions of QM either, as Stapp did it, by postulating yet another version of locality, or, as I suggested elsewhere, by appropriately rewriting QM predictions in terms of counterfactuals. But the original path leading to (H5) was not at all straightforward, and it required the highly suspicious assumption LOC2, contested by all subsequent critics.[6] On the other hand, here we have an intuitive, almost trivial line of argument, and the only help is taken from (EEC). If (EEC) were a reasonable assumption, Stapp wouldn't have to take the route he originally took in (1997).

But let us return to the current GHZ-type argument. Now it should be obvious that the entire argument collapses, as the chain leading to the sentence $Y_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (Y_1Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = p)$ is broken. But in spite of this let us continue analyzing the rest of the argument. So we will take it for granted that you can derive (4) from (2). Interestingly enough, the elimination of the next "intermediate" counterfactual condition in (3) is perfectly acceptable. Let us see why. We should be able to get from

(5) $Y_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (X_1Y_2Y_3 \;\Box\!\!\rightarrow (Y_1Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = p))$

to

[6] Elsewhere I have shown that both assumption LOC1 and LOC2, used in the Hardy-type argument, cannot be satisfied jointly under any intuitive reading for spatiotemporal counterfactuals. A similar point is made in (Finkelstein 1998).

(6) $Y_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (Y_1Y_2Y_3 \ \Box\!\!\rightarrow y_2y_3 = p)$.

So suppose that (5) holds. It means that the counterfactual $Y_1Y_2Y_3 \ \Box\!\!\rightarrow y_2y_3 = p$ must hold in all worlds $\langle X_1, Y_2, Y_3, \_, \_, \_\rangle$ which are closest to some of the worlds $\langle Y_1, X_2, X_3, \_, p, -p\rangle$ or $\langle Y_1, X_2, X_3, \_, -p, p\rangle$. But we have already noted that according to our rules of relative similarity, all worlds of the first type are equally similar with respect to all the latter worlds. Therefore it means that in order to make (5) true, the counterfactual $Y_1Y_2Y_3 \ \Box\!\!\rightarrow y_2y_3 = p$ must be valid *at all worlds* $\langle X_1, Y_2, Y_3, \_, \_, \_\rangle$. This, on the other hand, is true if $y_2y_3 = p$ is true at all worlds $\langle Y_1, Y_2, Y_3, \_, \_, \_\rangle$ closest to some of the worlds $\langle X_1, Y_2, Y_3, \_, \_, \_\rangle$. But now we see that *every* world of the type $\langle Y_1, Y_2, Y_3, \_, \_, \_\rangle$ is closest to *some* world of the type $\langle X_1, Y_2, Y_3, \_, \_, \_\rangle$: namely it is closest to the world in which Y_2 and Y_3 have respectively the same unchanged results, and X_1 has a result uniquely determined by (QM2). Therefore, we have finally shown that in order for (5) to be true, the sentence $y_2y_3 = p$ must be true at all worlds of the type $\langle Y_1, Y_2, Y_3, \_, \_, \_\rangle$. But if it is so, then obviously $y_2y_3 = p$ will be true at all worlds $\langle Y_1, Y_2, Y_3, \_, \_, \_\rangle$, and such that they are closest to some world fulfilling the antecedent of (6). Hence, in our model whenever (5) is true, (6) must be true as well.[7]

4. Partial realism to the rescue?

I will briefly sketch the next steps of the argument. Having proven (apparently) the validity of the sentence (6), Bedford and Stapp then combine it together with (1) in order to obtain the needed version of the implication (R1), namely

(C1) $X_2X_3 \wedge x_2x_3 = -p \Rightarrow (Y_2Y_3 \ \Box\!\!\rightarrow y_2y_3 = p)$

I admit that this move is a valid one, given that we accept yet another instance of the locality principle, this time referring only to the choice of measurements (this principle is guaranteed by the rule (CS2)):

[7] Of course we should stress that according to what has been said before, (5) is false in our model. But the validity of the move from (5) to (6) is not the trivial validity based on the assumption of the falsity of the antecedent (obviously we know that the material implication with a false antecedent must be true). In order to better see what we have argued for in the last paragraph, note that the entire argument for the validity of the transition from (5) to (6) does not depend on the particular form of the last consequent in (5) – instead of "$y_2y_3 = p$" we could have chosen any sentence φ.

(LOC″) $X_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (Y_1Y_2 \;\Box\!\!\rightarrow X_1)$

$Y_1X_2X_3 \wedge y_2y_3 = p \Rightarrow (Y_1Y_2 \;\Box\!\!\rightarrow Y_1)$

Using (LOC″) we can eliminate the reference to the measurement X_1 (or Y_1) from the antecedent of the counterfactual in (1) (or (6)). For example, we can argue for the validity of the following:

(6′) $Y_1X_2X_3 \wedge x_2x_3 = -p \Rightarrow (Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = p)$,

because, according to (LOC″), at the Y_2Y_3-world closest to the actual, Y_1 must hold, and therefore from (6) we know that $y_2y_3 = p$ must hold as well. In that way we can obtain both strict implications in (6′) and similarly constructed (1′) with the same consequent, and appealing to standard rules of logic we finally get (C1).

The same logical transformations (with the same flaw as we indicated before) lead from (QM3) and (QM4) to the counterfactual version of the step (R2):

(C2) $Y_2X_3 \wedge y_2x_3 = q \Rightarrow X_2Y_3 \;\Box\!\!\rightarrow x_2y_3 = q$

This time, the missing element which would allow us to get (C2) is the assumption of the reality of the objective value for Y_1. Summing up, we can say that Stapp is able to get his intermediary conclusions (C1) and (C2) only if he assumes the objective reality of values of two incompatible observables X_1 and Y_2 characterizing the first particle. But we will see in a moment that the very assumption that allows him to make these intermediary steps, will make another logical transition invalid, so the final conclusion will be unwarranted even on this additional supposition of “partial realism”.

Let us briefly go through the next steps of the argument. We will have to find a counterfactual counterpart of the step (R3). Obviously, we can’t have two results x_2 and y_2 defined simultaneously, but we can try to do this counterfactually. So let us assume that in reality two measurements X_2 and X_3 were performed with the results respectively $x_2 = m$ and $x_3 = n$. From (C1) one can obtain:

(7) $X_2 \wedge x_2 = m \wedge X_3 \wedge x_3 = n \Rightarrow (Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = -mn)$

In order to obtain the equation $y_3 = -mnr$ figuring in (R3), we must assume counterfactually that the result of the measurement of Y_2 was $y_2 = r$. Stapp inserts this supposition between the strict implication and the counterfactual conditional, in the form of yet another counterfactual:

(8) $X_2 \wedge x_2 = m \wedge X_3 \wedge x_3 = n \Rightarrow (Y_2 \wedge y_2 = r \wedge X_3 \;\Box\!\!\rightarrow (Y_2Y_3 \;\Box\!\!\rightarrow y_3 = -mnr))$

The crucial question is what justifies the move from (7) to (8). And here Stapp appeals to his second pattern of inference, which we have dubbed "the Principle of Addition of Irrelevant Conditions" (AIC). It essentially claims that when (7) is true, the following must be also true:

(9) $X_2 \wedge x_2 = m \wedge X_3 \wedge x_3 = n \Rightarrow (Y_2X_3 \;\Box\!\!\rightarrow (Y_2Y_3 \;\Box\!\!\rightarrow y_2y_3 = -mn))$

The transition from (9) to (8) is just a matter of using some unquestionable logical rules together with simple algebra, although its proper formalization can be tedious. But this is not a crux of the step from (7) to (8). The focal point is, of course, the principle (AIC). However, it appears that in this context the principle (AIC) is OK, although we should stress that generally the Lewis calculus of counterfactuals does not allow for such "insertions". I will not show right now that within our general semantic framework the move from (7) to (9) is valid, leaving it to the reader. In a nutshell, the only thing we must do is to show that all Y_2Y_3-worlds which are closest to some "actual" worlds (where actuality is of course understood according to what is said in the antecedent of the strict implication in (7)), will also be closest to these Y_2X_3-worlds which are closest to actuality.

The next few "logical twists and turns", with which I have no quarrel, lead from (9) via (C2) to the following chain of counterfactuals:

(10) $X_2 \wedge x_2 = m \wedge X_3 \Rightarrow (Y_2X_3 \;\Box\!\!\rightarrow (Y_2Y_3 \;\Box\!\!\rightarrow (X_2Y_3 \;\Box\!\!\rightarrow x_2 = -m)))$

And now the final goal seems to be within our reach: we should only get rid of these nasty intermediate counterfactuals in order to obtain something which looks almost like a contradiction:

(11) $X_2 \wedge x_2 = m \wedge X_3 \Rightarrow (X_2Y_3 \;\Box\!\!\rightarrow x_2 = -m)$

Actually, (11) contradicts our initial assumption of locality, which implies that changing counterfactually the observable X_3 for Y_3 should not change the result obtained in the measurement of X_2. So the only thing we should consider right now is how to justify the step from (10) to (11). In order to do this, we should not rely uncritically on the assumption (EEC), for we already know that this principle is not universally valid. Instead, we can again resort to our formal semantic model and ask if the truth of (10) can guarantee the truth of (11). Let us then symbolize the last counterfactual $X_2Y_3 \Box\!\!\rightarrow x_2 = -m$ in (10) by φ. If (10) is to be true, φ must be true in all Y_2Y_3-worlds which are closest to some Y_2X_3-worlds, which in turn must be closest to some of the worlds allowed by the antecedent of the strict implication.

Consider then first all worlds of the type $\langle \_, Y_2, X_3, \_, \_, \_\rangle$, and ask which of them are closest to some world of the type $\langle \_, X_2, X_3, \_, m, \_\rangle$, which is the type defined by the antecedent of the strict conditional in (10). In other words, we are asking what restrictions on possible Y_2X_3-worlds are put by the fact that we are considering only actual worlds satisfying the antecedent of the strict conditional. It appears that these restrictions are very mild. The only Y_2X_3-worlds for which we can find no world of the type $\langle \_, X_2, X_3, \_, m, \_\rangle$ such that the former is closest to the latter, are the worlds of the type $\langle X_1, Y_2, X_3, x_1, y_2, x_3\rangle$, where $x_1x_3 = m$. It is so because the worlds of the type $\langle X_1, X_2, X_3, x_1, m, x_3\rangle$, under supposition $x_1x_3 = m$ are forbidden by requirement (QM1). For all other Y_2X_3-worlds one can always find a world of the type $\langle \_, X_2, X_3, \_, m, \_\rangle$ to which it is closest, by putting in blanks the same elements as in the original one. This means that for (10) to be true, φ must be true in all Y_2Y_3-worlds such that there is an Y_2X_3-world fulfilling the above requirement, to which our Y_2Y_3-world is closest. But now it is obvious that we can find such a world of the type $\langle \_, Y_2, X_3, \_, \_, \_\rangle$ for all Y_2Y_3-worlds with no exception. To do this we should only replace the first blank by the measurement done in our initial Y_2Y_3-world, and insert in the next two blanks the results obtained in this world as well, completing the entire world by inserting a result for the measurement X_3 which agrees with quantum-mechanical predictions (QM1-4). Therefore it finally turns out that in order for (10) to be true, φ must be true at all Y_2Y_3-worlds, which allows us to cross out Y_2X_3 from (10) and to obtain:

(12) $X_2 \wedge x_2 = m \wedge X_3 \Rightarrow (Y_2Y_3 \Box\!\!\rightarrow (X_2Y_3 \Box\!\!\rightarrow x_2 = -m)))$.

The same reasoning can convince us that in order for (12) to be true, the sentence $x_2 = –m$ should be true at all X_2Y_3-worlds. Therefore, we can finally achieve (11) as our ultimate result.

To recapitulate: Stapp's entire reasoning appeared to be valid within our semantic model, with a notable exception of one crucial step, leading from (2) to (4); therefore the arguments fails. The purported non-locality of QM turns out to be unwarranted. However, one interesting thing remains to be analyzed. We remarked earlier that the minimal condition which would make the questioned move valid requires that both X_1 and Y_1 have definite values. It is interesting to ask whether this additional assumption would allow us to obtain the needed contradiction. Surprisingly, it appears that even this correction will not work, because in that case the transition from (10) to (11) ceases to be valid. In the remaining of the paper I will present a proof that this is the case.

Let us then construct another semantic model, with the built-in assumption that X_1 and Y_1 have already their values determined. This assumption implies that there will be substantially fewer possible worlds to consider; only those of the form $\langle X_1, \_, \_, a, \_, \_\rangle$ and $\langle Y_1, \_, \_, b, \_, \_\rangle$, with fixed a and b, will be allowed. Taking into consideration that the quantum-mechanical predictions (QM1–4) are still valid, we can easily calculate that the number of possible worlds in this case equals $2^5 – 8 = 24$ (an easier way of obtaining this result would be that the initial number 48 must be divided into 2, because we must reject all the possible worlds with results $–a$ and $–b$ for X_1 and Y_1 respectively).

Next, I will explicitly write down all possible worlds crucial for the evaluation of the counterfactuals in (10) and (11), giving them a convenient symbolization:

$w_0 = \langle X_1/Y_1, X_2, X_3, a/b, +1, –a\rangle$
$w_1 = \langle X_1/Y_1, X_2, X_3, a/b, –1, a\rangle$

$w_2 = \langle X_1/Y_1, Y_2, Y_3, a/b, +1, a\rangle$
$w_3 = \langle X_1/Y_1, Y_2, Y_3, a/b, –1, –a\rangle$

$w_4 = \langle X_1/Y_1, X_2, Y_3, a/b, +1, b\rangle$
$w_5 = \langle X_1/Y_1, X_2, Y_3, a/b, –1, –b\rangle$

$w_6 = \langle X_1/Y_1, Y_2, X_3, a/b, +1, b\rangle$
$w_7 = \langle X_1/Y_1, Y_2, X_3, a/b, –1, –b\rangle$

Suppose that the value m for X_2 figuring in (10) and (11) equals +1, and that $a = -b$. The entire argument looks exactly alike when we assume otherwise. Then we can easily show that (10) is true in our model (which strongly suggests that up to now all steps in the reasoning were valid, granting our assumption about objective values of X_1 and Y_1). The proof of this fact consists of the following statements, easily verifiable on the basis of (CS1):

$w_6 <_0 w_7$

$w_2 <_6 w_3$

$w_5 <_2 w_4$[8]

and we see that at the world w_5 indeed X_2 has value –1.

However, crossing out the first counterfactual antecedent from (10) is now unjustified! It is namely no longer true that every Y_2X_3-world is closest to some actual world. For example, as we noted, w_6 is closer to w_0 than w_7. There is no other world which would satisfy the antecedent of (10) and such that w_7 would be closest to it, for we excluded the possibility that X_1/Y_1 can have other values than a/b. So in order to make (10) true, $Y_2Y_3 \;\Box\!\!\rightarrow (X_2Y_3 \;\Box\!\!\rightarrow x_2 = -m)$ has to be true only at w_6, and because $w_2 <_6 w_3$, the sentence φ must be true at w_2. But observe that in order for (12) to be true, φ would have to be true at all Y_2Y_3-worlds, i.e. at w_2 and w_3. This cannot be guaranteed by the fact that φ is true at w_2.

The fact that we cannot proceed from (10) to (11) under the supposition that X_1 and Y_1 have determined values, is not a surprise at all. Note that we have just constructed a semantic model which has the following features: (i) it includes the supposition about determined values of X_1 and Y_1, (ii) it obeys the rules (CS1-3) of comparative similarity, which among other things assure that the locality assumption holds, and (iii) it agrees with quantum-mechanical predictions (QM1-4). If (11) were derivable from (10), which was shown to be true in our model, then (11) would have to be true as well. But (11) plainly contradicts the condition of locality, and therefore according to the stipulated rules (CS1-2) cannot be true in the above model. The very fact that we were able to construct the semantic model fulfilling all the conditions imposed by Stapp in his original proof, together with the assumption of the determinateness of the values of X_1 and Y_1, shows that no contradiction is derivable jointly from all these assumptions, and hence that even under "partial" realism the non-locality of

[8] When we assume that $a = b$, the following sentences form a proof of (10): $w_7 <_0 w_6$, $w_3 <_7 w_2$, $w_5 <_3 w_4$.

QM is unwarranted. For now, the only compelling argument that we know of for the non-locality must rely on the assumption of full-fledged realism of values for all observables involved.

Bibliography

Auletta, G.: 2000, *Foundations and Interpretations of Quantum Mechanics*, World Scientific, Singapore – New Jersey – London – Hong Kong.

Bedford, D and Stapp, H.P.: 1995, 'Bell Theorem in an Indeterministic Universe', *Synthese* 102, 139-164.

Bigaj, T.: 2002a, 'David Lewis's Logic of Counterfactuals and Quantum Physics', preprint.

Bigaj, T.: 2002b, 'Counterfactuals and Spatiotemporal Events', preprint.

Finkelstein, J.: 1998, 'Yet Another Comment on Nonlocal Character of Quantum Theory', preprint quant-ph/9801011

Finkelstein, J.: 1999, 'Space-time Counterfactuals', *Synthese* 119, 287-298.

Greenberger, D.; Horne, M.; Shimony, A., and Zeilinger, A.: 1990, 'Bell's Theorem without Inequalities', *American Journal of Physics*, 58, 1131-43.

Jarrett, J.P.: 1984, 'On the Physical Significance of the Locality Conditions in the Bell Arguments', *Nous* 18, 569-89.

Lewis, D.: 1973, *Counterfactuals*, Harvard Univ. Press, Cambridge (Mass.)

Lewis, D.: 1986, *Philosophical Papers Vol II*, Oxford.

Maudlin, T.: 1994, *Quantum Non-locality and Relativity*, Blackwell, Oxford.

Mermin, D.: 1998, ' Nonlocal Character of Quantum Theory?', *American Journal of Physics* 66(10), 920-924.

Placek, T.: 2000, *Is Nature Deterministic? A Branching Perspective on EPR Phenomena*, Jagiellonian University Press, Krakow.

Stapp, H.: 1971, 'S-matrix Interpretation of Quantum Theory', *Physical Review* 3D, 1303-20.

Stapp, H.: 1989, 'Quantum Nonlocality and the Description of Nature' in J. Cushing and E. McMullin (eds.) *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem*, University of Notre Dame Press, Notre Dam e.

Stapp, H.: 1991, 'Quantum Measurement and the Mind-Brain Connection' in P. Lahti and P. Mittelstaedt (eds.), *Symposium on the Foundations of Modern Physics 1990*, World Scientific, Singapore – New Jersey – London – Hong Kong.

Stapp, H.: 1997, 'Nonlocal Character of Quantum Theory', *Am. J. Phys*. 65, 300-304.

Stapp, H.: 1999, 'Comments on 'Non-locality, Counterfactuals, and Quantum Mechanics'', *Physical Review A* 60(3), 2595-2598.

Redhead, M.: 1987, *Non-locality, Incompleteness and Realism*, Oxford Univ. Press, Oxford.

Redhead, M., Butterfield, J. and Clifton R.: 1991, 'Nonlocal Influences and Possible Worlds', *Brit. J. Phil. Sci.* 41, 5-58.

Redhead, M., La Riviere, P.: 1997, 'The Relativistic EPR Argument' in R.S. Cohen, M. Horne and J. Stachel (eds.), *Potentiality, Entanglement and Passion-at-a-Distance*, Dordrecht.

Unruh, W.: 1999, 'Nonlocality, Counterfactuals, and Quantum Mechanics', *Phys. Rev.* A, 59, 126-130.

van Fraassen, B.: 1991, *Quantum Mechanics. An Empiricist View*, Clarendon Press, Oxford.