

ARTICLES

Strong versions of Bell's theorem

Henry P. Stapp

Theoretical Physics Group, Physics Division, Lawrence Berkeley Laboratory, 1 Cyclotron Road, Berkeley, California 94720

(Received 1 June 1993)

Technical aspects of a recently constructed strong version of Bell's theorem are discussed. The theorem assumes neither hidden variables nor factorization, and neither determinism nor counterfactual definiteness. It deals directly with logical connections. Hence its relationship with modal logic needs to be described. It is shown that the proof can be embedded in an orthodox modal logic, and hence its compatibility with modal logic assured, but that this embedding weakens the theorem by introducing as added assumptions the conventionalities of the particular modal logic that is adopted. This weakening is avoided in the recent proof by using directly the set-theoretic conditions entailed by the locality assumption.

PACS number(s): 03.65.Bz

I. POSTULATE

A strong version of Bell's theorem [1] has recently been proved. It involves propositions such as the following statement S : "If measurement M were to be performed and its outcome were to be O_1 then if, *instead of* M , N were to be performed, its outcome would be O_2 ."

Valid statements of this kind arise naturally in physics, as implications of particular theories. For example, suppose the two alternative possible measurements M and N are obtained by placing either a device D_M or a device D_N into a beam of charged particles of known charge and mass. Suppose classical electromagnetic theory entails that each of the individual particles in the beam *must* land at the same place on a detecting screen independently of which of the two alternative possible devices is used. Then, on the basis of classical electromagnetic theory, one could affirm the validity of statement S with O_1 the same as O_2 . Moreover, this statement would remain valid in the context of a theory in which only statistical information is available about which of the possible trajectories a given particle follows.

Statements like S are appropriate tools for formulating a certain concept of "no faster-than-light influences." If the two alternative possible measurements M and N differ only by a randomly determined choice of whether the device D_{AM} or the device D_{AN} is used in region R_A , so that, in either case M or N , exactly the same setup and device, $D_{BM} \equiv D_{BN} \equiv D_B$, is used in the spacelike-separated region R_B , and if, moreover, the outcomes O_1 and O_2 refer only to what appears in region R_B , then statement S , with $O_1 \equiv O_2$, becomes an expression of the idea that there is no faster-than-light influence of any kind. It says that, for a single fixed experimental setup in region R_B , the outcome appearing there must be independent of which of the two devices D_{AM} or D_{AN} is chosen in the far-away region R_A (at the same instant of time in

some frame of reference).

The clause "If, *instead of* M , N . . ." is called a counterfactual conditional by logicians. Formal rules have been devised for the use of such conditionals. These rules are constructed so as to allow the meanings of the relevant statements to be retained, insofar as these meanings can be agreed upon and consistently maintained. Probably the "most orthodox" formulation of these rules is the one proposed by Lewis [2].

The strong Bell's theorem mentioned above was proved in Ref. [3]. A somewhat similar argument was given in Ref. [4]. The latter argument was technically more complicated because it was based on the Greenberger-Horne-Zeilinger experiment, which involves experiments in three regions, instead of the Hardy experiment, which involves experiments in only two regions. Moreover, it was formulated within the context of Lewis's formalism. It is possible to interpret Ref. [3] as simply a Hardy version of the argument in Ref. [4], and hence to construe the counterfactual conditional appearing there as Lewis counterfactual conditionals. It will be useful to consider that interpretation first.

All steps *save one* in this Lewis version of the argument of Ref. [3] are strictly justifiable within the Lewis framework: that was shown in Ref. [4]. In Ref. [4] the step corresponding to the single exception was justified by a postulated special rule of inference called "elimination of eliminated conditions." This postulate is called here EEC.

What is this postulate EEC? Starting from the assumptions of the theorem (random choices of measurement, unique outcomes of measurements, validity of the predictions of quantum theory, and absence of faster-than-light influences) and using only logical steps justified within the Lewis framework, one arrives at a long proposition that translated into words says the following.

"If A_1 is performed in R_A , and B_1 is performed in R_B , and if the outcome in R_A is 'yes,' then:

[if, instead of A_1 and B_1 , A_1 and B_2 are performed then
 [if, instead of A_1 and B_2 , A_2 and B_2 are performed then
 [if, instead of A_2 and B_2 , A_2 and B_1 are performed then
 [if, instead of A_2 and B_1 , A_1 and B_1 are performed then
 [the outcome in R_B would be 'no']]]]]]."

The postulated rule of inference EEC is restricted to situations in which (1) the choices that determine which measurements are performed in the two regions are treated as independent random variables, and (2) no reference is made in the long proposition to the *outcomes* of any of the (randomly chosen) intermediate sets of measurements. Under these conditions EEC asserts that the exactly countermanded conditions can be ignored, and the long proposition reduced to the shorter statement:

"If A_1 is performed in R_A and B_1 is performed in R_B , and if the outcome in R_A is 'yes,' then if A_1 is performed in R_A and B_1 is performed in R_B , the outcome in R_B would be 'no.'"

The validity of this step is justified by referring to the physical meanings of statements involved. If no reference at all is made to the outcome of a certain measurement N , which, moreover, is selected by a purely random decision that is unrelated to anything that comes before, then the condition

"If, instead of M , N is performed"

followed immediately by the assertion that then

"If, *instead* of N , P is performed"

is physically equivalent to

"If, instead of M , P is performed."

This is because the clause "*instead* of N " exactly cancels, in this case, the earlier supposition that " N is performed."

To complete this proof it should be shown that EEC is compatible with Lewis's rules. This is done in Sec. IV. First, the Lewis theory must be described, and its incompleteness noted.

II. LEWIS'S THEORY

Lewis's rules are defined over a set W of possible worlds. These rules are based on a concept of the "closeness" of possible worlds, where closeness is related to the laws of nature. Lewis's rules of closeness, as they apply to our indeterministic case, can be formulated as follows.

Each possible world w is defined on a corresponding spacelike surface $\sigma(w)$. This surface separates space-time

into two open sets, called the future and past of w . A world w_1 is *later* than a world w_2 if and only if some point in the past of w_1 lies in the future of w_2 , but no point in the past of w_2 lies in the future of w_1 . Each possible world w can evolve into later possible worlds. Let w_a and w_b be two possible worlds. Then $W(w_a, w_b)$ is defined to be the set of possible worlds w such that (1) w can evolve into w_a without violating any strict law of nature; and (2) w can evolve into w_b without violating any strict law of nature. Then w is said to be *closer* to w' than to w'' if the union of the pasts of the worlds in $W(w, w')$ is a proper subset of the union of pasts of the worlds in $W(w, w'')$.

This rule of closeness applies in an indeterministic universe in which, however, several constraints are rigidly enforced. In our case, these rigid constraints are the (100% certain) quantum predictions and the (strictly enforced) demand that there be no faster-than-light influence of any kind. The rules of closeness are used to determine, within the formalism, the truth or falsity of statements involving counterfactual conditions.

Consider, for example, our statement S . Symbolically, it is written

$$(M \wedge O_1) \Rightarrow (N \square \rightarrow O_2),$$

where \Rightarrow represents "implies" (the strict conditional) and $N \square \rightarrow$ represents "if, *instead* of M , N is performed then" (the counterfactual conditional). Statement S is asserted to be true, in the Lewis framework, if and only if *each* world w satisfying $(M \wedge O_1)$ is closer to *some* world w' satisfying $(N \wedge O_2)$ than to *any* world w'' satisfying $(N \wedge -O_2)$. (The symbol \wedge represents conjunction, and the minus sign represents negation.)

To see how the theory works, let us see how it validates the statement S given above in the case $M = A_1 \wedge B_1$, $N = A_2 \wedge B_1$, and $O_1 = O_2 =$ "no" in R_B , under the physical condition that there can be no faster-than-light influence of any kind.

Let w be any world in $\{M \wedge O_1\}$, which is the set of worlds satisfying the condition $(M \wedge O_1) = (A_1 \wedge B_1 \wedge \text{"no" in } R_B)$. Let w'' be any world in $\{N \wedge -O_2\}$, which is the set of worlds satisfying $(N \wedge -O_2) = (A_2 \wedge B_1 \wedge \text{"yes" in } R_B)$. Let $W(w, w'')$ be the set of worlds w_1 such that (1) w_1 can evolve into w without violating any strict law of nature, and (2) w_1 can evolve into w'' without violating any strict law of nature. Such worlds w_1 may exist, because the decision in R_A between A_1 and A_2 is a chance event, which can go either way without violating any strict law of nature, and the selection in R_B of the outcome "yes" or "no" is likewise able to go either way

without violating any of our strict laws. However, no such w_1 can contain in its *past* either the decision point p in R_A between A_1 and A_2 , or the decision point q in R_B between the outcomes “yes” and “no.” For if either of these decisions is already fixed in w_1 , then the two conditions (1) and (2) on w_1 cannot *both* be met: at most (1) *or* (2) can be satisfied. Thus the union of the pasts of the worlds in $W(w, w'')$ is confined to the past of w minus $\{V^+(p) \cup V^+(q)\}$, where $V^+(x)$ is the closed forward light cone with apex at x , and \cup means “union.” Consider now the world w . The physical condition that there be “no faster-than-light influence of any kind” means that any evolution is at least *allowed* to be independent of the choice of measurement made at any spacelike-separated point. Thus there is a world w' in $\{N \wedge O_2\} = \{A_2 \wedge B_1 \wedge \text{“no” in } R_B\}$ that is the same as w outside the light cone $V^+(p)$. Consider next the set of worlds $W(w, w')$. The past of any world in $W(w, w')$ is confined to the past of w minus $V^+(p)$, because the decision at p is A_1 in w , but is A_2 in w' . However, the decision point q lies in the past of both w and w' , because the decision there is “no” for both w and w' . Assume that if a world can evolve into a later one then there will be, in the set of all possible worlds, sequences corresponding to all possible ways in which the past of the earlier world can grow into the past of the later one. Then the union of the pasts of the worlds in $W(w, w')$ will be precisely the past of w minus $V^+(p)$, whereas the union of the pasts of the worlds in $W(w, w'')$ is confined to this set minus $V^+(q)$. Thus w is closer to w' than to any w'' in $\{N \wedge -O_2\}$, and the Lewis condition for the truth of statement S is satisfied.

This shows that Lewis’s theory *works* in this case: it yields the conclusion demanded by intuition.

III. INCOMPLETENESS OF LEWIS’S RULE OF CLOSENESS

The Lewis rule of closeness given above is formulated not in terms of the absolute distances between worlds, but rather in terms of the *relative* closeness of two worlds to a

third one. Moreover, this rule of closeness has only the “if” condition: the “only if” part is not included. Thus further rules are needed if the formalism is to provide a definite answer, true or false, to every statement that contains a counterfactual conditional. This opens up the issue of fine tuning, i.e., the problem of resolving those questions of closeness left unanswered by the primary rule of closeness given above. Lewis gives several lower-level rules that can resolve *some* of the issues of closeness not resolved by the primary rule, but they do not apply in the present context, where every strict law of nature is rigidly enforced. Lewis gives one further rule that does apply. This is the *centering* rule: any world is closer to itself than to any other world. This centering rule suggests that we are dealing with a metric space.

A simple situation in which the Lewis rules are mute is the case $M = (A_1 \wedge B_1)$ and $N = (A_2 \wedge B_2)$. Then the worlds w_1 in $W(w, w')$ and w_2 in $W(w, w'')$ are both blocked by the same condition: neither decision point between measurements can lie in the past of either w_1 or w_2 . Thus there is a “tie,” and the Lewis rules of closeness are insufficiently discriminatory to allow any conclusion to be drawn: more detailed rules of closeness are needed for the Lewis truth rule to yield a result.

IV. COMPATIBILITY OF EEC WITH THE LEWIS RULES

Because of the incompleteness of the Lewis rule of closeness, the Lewis rules are not sufficiently complete to permit the validity of the postulate EEC to be either confirmed or rejected: that is why the postulate was introduced. But if a new postulate is introduced then its *compatibility* with the old ones should be verified.

Consider the statement

$$S_1 \equiv (P \square \rightarrow O_2) .$$

The general Lewis truth rule asserts that the set of worlds in which S_1 is true is

$$\{S_1\} \equiv \{P \square \rightarrow O_2\} = \{w : |w - w'| < |w - w''| \text{ for some } w' \text{ in } \{P \wedge O_2\} \text{ and every } w'' \text{ in } \{P \wedge -O_2\}\} . \tag{4.1}$$

Here $\{x : C\}$ is the set of x such that condition C is satisfied, and $|w - w'| < |w - w''|$ means w is closer to w' than to w'' .

Consider next the compound statement

$$S_2 \equiv (N \square \rightarrow S_1) ,$$

where S_1 is the statement defined previously. The general Lewis truth rule asserts that the set of worlds in which S_2 is true is

$$\{S_2\} \equiv \{N \square \rightarrow S_1\} = \{w : |w - w'| < |w - w''| \text{ for some } w' \text{ in } \{N\} \cap \{S_1\} \text{ and every } w'' \text{ in } \{N\} \cap \{-S_1\}\} . \tag{4.2}$$

The symbol \cap means "intersection." Note that, by virtue of the centering rule, if N were the null condition, i.e., if $\{N\} = W$, then $\{S_2\} = \{S_1\}$.

The question at issue here is whether it is compatible with the general truth rule and the centering rule to take $\{S_2\} = \{S_1\}$, i.e., to take $\{N \square \rightarrow (P \square \rightarrow O_2)\} = \{P \square \rightarrow O_2\}$ for all N . This question is equivalent to that of whether it is possible to take $\{S_2\}$ to be independent of N .

To show that this is possible, let closeness be defined in a metric space (X, Y) , where X and Y are orthogonal. Suppose that each world w in W maps to a unique point $(x(w), y(w))$ in (X, Y) , and that each point (x, y) in (X, Y) maps to a unique world $w(x, y)$ in W . Then the centering rule is satisfied. Let $\hat{x}(M)$ be a function of the measurements. Suppose that measurement M is performed in world w if and only if $x(w) = \hat{x}(M)$. Then the set of possible worlds in which measurement M is performed and outcome O_i^M occurs can be written as

$$\{M \wedge O_i^M\} = \{w : x(w) = \hat{x}(M); y(w) \in \{O_i^M\}\},$$

where \in means "is an element of," and $\{O_i^M\}$ is a subset of Y space. [Each M is a set of local measurements, with one local measurement (possibly the null measurement) for each region, and each O_i^M is a set of local outcomes, with one local outcome (possibly the null outcome) for each of the local measurements in M .]

The set of worlds where $(P \wedge O_2^P) \equiv (P \wedge O_2^P)$ is satisfied is

$$\{P \wedge O_2^P\} = \{w : x(w) = \hat{x}(P); y(w) \in \{O_2^P\}\}.$$

Similarly,

$$\{P \wedge -O_2^P\} = \{w : x(w) = \hat{x}(P); y(w) \in \{-O_2^P\}\},$$

where $\{-O_2^P\}$ represents the complement of $\{O_2^P\}$ in Y space. Thus, by virtue of (4.1) and the orthogonality of X space to Y space,

$$\{S_1\} \equiv \{P \square \rightarrow O_2^P\} = \{w : y(w) \in \{O_2^P\}\}$$

and

$$\{-S_1\} = \{w : y(w) \in \{-O_2^P\}\}.$$

Thus, by intersection,

$$\{N\} \cap \{S_1\} = \{w : x(w) = \hat{x}(N); y(w) \in \{O_2^P\}\}$$

and

$$\{N\} \cap \{-S_1\} = \{w : x(w) = \hat{x}(N); y(w) \in \{-O_2^P\}\}.$$

Thus, by virtue of (4.2) and the orthogonality of X space to Y space,

$$\{S_2\} = \{w : y(w) \in \{O_2^P\}\}.$$

Hence

$$\{S_2\} = \{S_1\}.$$

In this model the statement

$$(M \wedge O_1^M) \equiv (N \square \rightarrow O_2^N)$$

is equivalent to

$$\{O_1^M\} \subset \{O_2^N\},$$

where \subset means "is a subset of." The statement

$$(N \wedge O_2^N) \equiv (P \square \rightarrow O_3^P)$$

is equivalent to

$$\{O_2^N\} \subset \{O_3^P\}.$$

Thus the two statements together say that

$$\{O_1^M\} \subset \{O_2^N\} \subset \{O_3^P\}.$$

This implies that

$$\{O_1^M\} \subset \{O_3^P\},$$

which is equivalent to

$$(M \wedge O_1^M) \equiv (P \square \rightarrow O_3^P).$$

The same line of argument validates EEC, and hence confirms the compatibility of EEC with the general Lewis framework. Arranging for the compatibility with also the Lewis rules of closeness and our locality condition poses no problem, but if one tries to impose, moreover, compatibility with the quantum predictions then a contradiction of course ensues.

V. IMPLEMENTATION OF THE "MUST" CONDITIONS

The Lewis framework normally validates statements that are not consequences of merely the laws of nature alone. One consequence of this fact is that we were able to validate statement S , which appears to express the strong locality condition that outcomes *must* be independent of spacelike-separated choices, from the weak locality condition that evolutions are *allowed* to be independent of spacelike-separated choices. Thus statement S , interpreted according to the Lewis rules, fails to carry the full logical content of the strong locality condition assumed in the statement of the theorem.

To exhibit this essential deficiency of the (unelaborated) Lewis counterfactuals in the present context, let us review how they worked in the construction described in Sec. II. The key step was the implementation of the weak locality condition. It allowed us to assert that for each world w in $\{A_1 \wedge B_1 \wedge \text{"no" in } R_B\}$ there was *some* world w' in $\{A_2 \wedge B_1 \wedge \text{"no" in } R_B\}$ that is the same as w outside the light cone $V^+(p)$. According to this argument, there might be only *one* such w' , but billions of possibilities in which the switching of A_1 to A_2 leads to a change of "no" to "yes" in R_B . But the existence of this single world w' with outcome "no" in R_A among billions of contrary possibilities is a very weak condition: the strong conclusion derived within the Lewis framework from the tiny result rests heavily on conventional rules, as contrasted to the strict laws of nature.

It is desirable from a certain point of view to derive our result within the Lewis framework. This framework is

probably the standard theory of counterfactuals; hence deriving the result within the strictures imposed by that theory lends credibility to the conclusion. But three important and related points must be borne in mind: (1) the Lewis framework is a collection of many theories that differ by the fine tuning of the rules of closeness, and we are free to choose any one that fits the physical requirements; (2) the conclusions drawn from a use of the formalism reflect in large measure certain conventional rules, as contrasted to strict laws of nature; and (3) the above proof within the Lewis framework uses only weak locality, not the strong locality condition assumed in the statement of the theorem.

To tailor the Lewis theory of counterfactuals to the problem at hand, it is necessary to impose the following condition: each world in $\{M \wedge O_i^M\}$ is *closer* to some world in $\{N \wedge O_j^N\}$ than to any world in $\{N \wedge -O_j^N\}$ if and only if the following condition holds.

“If under the condition that M were to be performed the outcome O_i^M were to occur then under the condition that, *instead* of M , N were to be performed, the outcome O_j^N *must*, by virtue of the (assumed) strict laws of nature, occur.”

Such a definition of “closeness” would make the Lewis formalism relevant to the theorem being proved. The specification is compatible with Lewis’s rules of closeness, but is much more restrictive. I shall call the consequent “under the condition . . . O_j^N *must* . . . occur” a *must* conditional.

The proof in Ref. [3] does not explicitly use Lewis counterfactuals: they are never mentioned. It thereby avoids the dependence on conventional definitions of “closeness.” It also avoids the need for arguments like the one exhibited in Sec. II. It circumvents those problems by exploiting *directly* the assumed strong locality condition, as expressed in terms of *must* conditionals. On the other hand, the proof can be embedded within the general Lewis framework by imposing the strong condition of closeness specified above.

The proof in Ref. [3] proceeds (working now from left to right) by combining one consequence of strong locality with one consequence of quantum theory to obtain the conclusion that *each* world w in $\{A_1 \wedge B_1 \wedge \text{“yes” in } R_A\}$ *must*, by virtue of our two strict laws of nature, become *some* world in $\{A_1 \wedge B_2 \wedge \text{“yes” in } R_B\}$ if the choice leading to B_1 is replaced by a choice leading to B_2 . There is a similar strict condition that *each* world w' in $\{A_1 \wedge B_2 \wedge \text{“yes” in } R_B\}$ *must* become *some* world in $\{A_2 \wedge B_2 \wedge \text{“no” in } R_A\}$ if the choice leading to A_1 is replaced by a choice leading to A_2 .

The *must* conditionals used here are defined over a set of worlds in which there is a set of disjoint alternative possible universes, each labeled by an alternative possible choice of the combined set of measurements. Initially, no connection whatever is imposed between these alternative possible universes: every possibility is independently allowed in each universe; apart, of course, from the specification that some particular combination of measurements is performed in each universe. No condition within a universe, or between two universes, is allowed unless it is *demand*ed by the strict laws of nature. These

strict laws are expressed by *must* conditionals of the form specified above.

These *must* conditionals are similar linguistically to Lewis’s counterfactual *would* conditionals. But they are much stronger. They do not, initially or basically, refer to the idea of closeness of worlds, but express, instead, a condition of set-theoretic inclusion under an “instead of” mapping that follows from the (assumed) strict laws of nature alone.

This *must* conditional can be represented symbolically by

$$\{M \wedge O_1^M\}_N \subset \{N \wedge O_2^N\}.$$

It says that *each* world in $\{M \wedge O_1^M\}$ is mapped by the “instead of” mapping into *some* world in $\{N \wedge O_2^N\}$.

Suppose an experimenter has three measuring devices M , N , and P , and that he has measured with device M and outcome O_1^M has occurred. Suppose he knows, on the basis of a theory, that if measuring with M were to give O_1^M then measuring *instead* with N *must* give O_2^N , and also that if measuring with N were to give O_2^N , then measuring *instead* with P *must* give O_3^P . Then he knows, on the basis of this theory, that if he had measured with P instead of M then result O_3^P necessarily would have occurred. The fact that this strong conclusion does not follow from the logically much weaker Lewis counterfactual conditionals, without any stipulations, is irrelevant. On the other hand, if one wishes to imbed these strong *must* conditionals into the Lewis framework one can do so by using the model described in Sec. IV.

If, in the above example, $P = M$ and $O_3^P = -O_1^M$ then one can conclude that the (assumed) theory is self-contradictory, provided the theory also entails that cases in which measurement with M gives an outcome O_1^M *must* be allowed. This is the kind of logical contradiction that occurs if a theory imposes jointly the conditions that the predictions of quantum theory be valid and that there be no faster-than-light influence of any kind.

VI. CONCLUSION

Two different versions of the argument of Ref. [3] have been discussed here. The first is essentially the proof given in Ref. [4], simplified to the Hardy case. This version is based on the standard Lewis theory of counterfactuals and involves postulating EEC.

The possible challenge to the postulation of EEC has been met by demonstrating that EEC is compatible with the Lewis rules. Thus it is permissible to postulate it, within the framework, without committing any logical error. However, a detailed discussion of the Lewis counterfactual version of the argument has allowed us to pinpoint a serious deficiency of that type of approach to the problem under consideration here. The conclusions drawn from a Lewis-type analysis rest heavily upon certain conventional rules pertaining to the notion of the “closeness of worlds.” Consequently, the conclusions obtained from a Lewis-based formulation of the argument would not necessarily follow exclusively from the two assumed strict laws of nature themselves, and the other two

assumptions set forth in the statement of the theorem: strong special conditions on the rules of closeness are needed if the consequences of applying the Lewis theory are to be strict consequences of the assumptions of the theorem alone.

The proof given in Ref. [3] does not use Lewis counterfactual conditionals: they are never mentioned in that paper. It uses, instead, the assumed strict laws of nature themselves, and in particular the statements of strong locality and quantum laws expressed in terms of *must* conditionals. Thus the dependence upon the notion of closeness of worlds, and upon the attendant conventional rules of closeness, is completely avoided. On the other hand,

the argument can be placed within the general Lewis framework, if appropriate rules of closeness are introduced. These rules are restricted by the strong condition that the consequences of applying these rules should be consequences of the assumptions of the theorem alone.

ACKNOWLEDGMENTS

This work was supported by the Director, Office of Energy Research, Office of High Energy and Nuclear Physics, Division of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

[1] J. S. Bell, *Physics* **1**, 95 (1964).

[2] D. Lewis, *Counterfactuals* (Blackwell, Oxford, 1976); *Philosophical Paper* (Oxford University Press, Oxford, 1986), Vol. 2.

[3] H. P. Stapp, *Phys. Rev. A* **46**, 6860 (1992).

[4] D. Bedford and H. P. Stapp, Lawrence Berkeley Laboratory Report No. LBL-29836 (unpublished); and Synthese (to be published).