Foundations of Physics, Vol. 29, No. 5, 1999

Time-Symmetrized Counterfactuals in Quantum Theory[1]

Lev Vaidman[2]

Received March 18, 1999

Counterfactuals in quantum theory are briefly reviewed and it is argued that they are very different from counterfactuals considered in the general philosophical literature. The issue of time symmetry of quantum counterfactuals is considered and a novel time-symmetric definition of quantum counterfactuals is proposed. This definition is applied for analyzing several controversies related to quantum counterfactuals.

1. COUNTERFACTUALS IN THE CONTEXT OF QUANTUM THEORY

There are very many philosophical discussions on the concept of counterfactuals and, especially, on the time's arrow in counterfactuals. There is also a considerable literature on counterfactuals in quantum theory. In order to be a helpful tool in quantum theory, counterfactuals have to be rigorously defined. Unfortunately, the concept of counterfactuals is vague[3] and this leads to several controversies. I, however, believe that since quantum counterfactuals appear in a much narrower context than in general discussions on counterfactuals, they can be defined unambiguously. I briefly review counterfactuals in quantum theory and propose a rigorous definition which can clarify several issues, in particular, those related to the time symmetry of quantum counterfactuals.

[1] The present paper, which has been in the public domain as a preprint,[(1)] is critically analyzed in a forthcoming paper by Kastner.[(2)]

[2] School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel-Aviv 69978, Israel.

[3] "Counterfactuals are infected with vagueness, as everybody agrees" (Ref. 3, p. 34).

0015-9018/99/0500-0755$16.00/0

A general form of a counterfactual is as follows.

(i) *If it were that $\mathcal{A}$, then it would be that $\mathcal{B}$.*

The basic approach to analyzing counterfactuals is to consider the *actual* world, the world that we know, in which $\mathcal{A}$ is in general not true, and a counterfactual world, *closest* to the actual world, in which $\mathcal{A}$ is true. The truth of the counterfactual (i) depends on the truth of $\mathcal{B}$ in this counterfactual world.

There is a general philosophical trend to consider counterfactuals to be asymmetric in time. Even Bennett,[4] who was challenging this claim in 1984, reversed his position (as I learned from private correspondence). In the most influential paper on this subject, Lewis (Ref. 3, p. 37) writes,

> I believe that indeterminism is neither necessary nor sufficient for the asymmetries I am discussing. Therefore I shall ignore the possibility of indeterminism in the rest of this paper, and see how the asymmetries might arise even under strict determinism.

In contrast to this opinion, I believe that the indeterminism is crucial for allowing nontrivial time-symmetric counterfactuals and that Lewis's and other general philosophical analyses are irrelevant for the issue of counterfactuals in quantum theory. The key questions in these analyses are related to $\mathcal{A}$: Why $\mathcal{A}$, if in the actual world $\mathcal{A}$ is not true? Do we need a "miracle" (i.e., breaking the laws of physics) for $\mathcal{A}$? Does $\mathcal{A}$ come by itself, or it is accompanied by other changes? In contrast, in the context of quantum theory there are no important questions related to $\mathcal{A}$. In some cases, $\mathcal{A}$ is related to an external entity which might vary freely by fiat; in other cases, the indeterminism of the theory allows different $\mathcal{A}$ without need for "miracles"—the main topic of discussion on counterfactuals in general philosophy.

The main source of vagueness in counterfactuals is in the definition of a counterfactual world *closest* to the actual world. Clearly, it differs in $\mathcal{A}$. In a deterministic world, other differences are also required: a "miracle" for $\mathcal{A}$ to happen, etc. There is no rigorous specification of aspects of a counterfactual world which are fixed to be identical to those of the actual world. The definition of such specification is missing in most discussions on quantum counterfactuals too. The main result of this work is a proposal for such definition. The most important feature of this definition is that it is also applicable for time-symmetric situations.

In the literature on quantum theory there are two main (different) concepts named "counterfactuals." Quantum counterfactuals of the first type are events which did not happen in our world but somehow influenced it. To present this concept, let me quote Penrose (Ref. 5, p. 240):

> What is particularly curious about quantum theory is that there can be actual physical effects arising from what philosophers refer to as *counterfactuals*—that is, things that might have happened, although they did not happened.

In particular, Penrose's quotation relates to interaction-free measurements[6] in which a location of supersensitive mine, which explodes if anything "touches" it, can be found without an explosion. The counterfactual here is the explosion which could have happened but didn't. What allows such counterfactuals without miracles is the *indeterminism* of the quantum theory (with collapse). In a noncollapse deterministic interpretation such as the Many-Worlds Interpretation of quantum theory,[7,8] the explanation is different (and, in my option, is particularly clear). The counterfactuals are "actual" in other worlds (Ref. 9, p. 275). Thus, in the situations considered by Penrose, "things" did happen in the physical universe (the union of all worlds) and thus their effect on some other facts in the physical universe is not so surprising.[10]

The counterfactuals of the first type are certainly helpful: they provide deeper explanations of many peculiar quantum phenomena. For example, we can understand why there is an "interaction-free" measurement which can ascertain that in a certain location there is a supersensitive mine, but there is no "interaction-free" measurement ascertaining that in a certain location there is *no* supersensitive mine: in the latter there is no counterfactual world (such as the world with the explosion in the previous case) different from the actual one. However, quantum counterfactuals of the first type cannot be brought to the general form (i) and they are not the main topic of this paper.

Quantum counterfactuals of the second type are statements in form (i) related to a close quantum system. $\mathscr{A}$ defines which experiments are performed on this system by an external observer and $\mathscr{B}$ is related to the results of these experiments. The decision of the observer which experiments to perform is assumed to be independent on the state of the quantum system under investigation. One can freely change everything outside the quantum system in question. This aspect represents a crucial difference between quantum counterfactuals and the counterfactuals in the general philosophical literature where $\mathscr{A}$ is related to the whole world.

Most examples of quantum counterfactuals discussed in the literature are in the context of EPR-Bell-type experiments.[11–13] Bedford and Stapp[14] presented an analysis of a Bell-type argument in the formal language of the Lewis theory of counterfactuals.[15] Most recently the work of Stapp,[16] based on the Hardy-type experiment,[17] was followed by intensive polemic.[13–18] A typical example is a consideration of an array of incompatible measurements on a composite system in an entangled state. Various conclusions are derived from statements about the results of these

measurements. Since these measurements are incompatible, they cannot all be performed together, so it must be that at least some of them were not actually performed. This is why they are called counterfactual statements.

2. DEFINITION OF TIME-SYMMETRIZED QUANTUM COUNTERFACTUALS

Quantum counterfactuals are usually explicitly asymmetric in time. The asymmetry is neither in $\mathcal{A}$ nor in $\mathcal{B}$; both are about the *present* time. The asymmetry is in the description of the actual world. The *past* but not the *future* of a system is given.

My purpose here is to avoid the asymmetry in time and to allow both the past and the future of counterfactual worlds to be fixed. However, it seems that $\mathcal{A}$ changes the future and therefore the future cannot be kept fixed. Indeed, the complete description of a quantum system is given by its quantum state, and the choice of measurements, described by $\mathcal{A}$, changes the future quantum state to be one of the eigenstate of the measured variable. Therefore, we cannot hold fixed the quantum state of the system in the future.

The way to overcome this difficulty is not to use a quantum state as the description of a physical system. For solving the current problem we can consider the quantum state only as a mathematical tool for calculating the probabilities of the results of measurements, and not as a description of the "reality" of a quantum system. Indeed, counterfactual statements are related to our experience which is connected to a quantum system through results of experiments. Therefore, we can define counterfactuals in terms of results of experiments without entering the issue of the "reality" of a quantum system. The advantage of this pragmatic approach is that it is universal: it fits all interpretations of quantum theory. Thus, my proposal for defining counterfactuals in quantum theory is as follows.

> (ii) *If measurement $\mathcal{M}'$ instead of measurement $\mathcal{M}$ has been performed on a system S, then the outcome of $\mathcal{M}'$ would have property $\mathcal{P}$. The results of all other measurements performed on system S are fixed.*

$\mathcal{M}$ and $\mathcal{M}'$ consist, in general, of measurements of several observables performed at space–time points P_i. The property $\mathcal{P}$ is a certain relation between the results of these measurements or a probability for such relation to happen.

What makes my definition different and rigorous is the clarification of what is fixed. Usually, this is not spelled out and it is tacitly assumed that

the quantum state of the system prior to the times of the space–time points P_i is fixed.

In usual time-asymmetric situations, in which the past relative to P_i exists but the future does not, the counterfactuals according to my definition are identical to those in the usual approach. Indeed, the results of all measurements in the past define the quantum state uniquely. No controversies appear in such cases: the past of the counterfactual worlds is fixed to be the past of the actual world. The problems arise when there is some information about the future of a system, sometimes, "future" only according to a particular Lorentz frame. (For systems consisting of spatially separated parts, the "past" and "future" depend on the choice of the Lorentz frame.) Following the principle that only the past is fixed and bringing together "true" counterfactuals from various Lorentz frames frequently leads to paradoxes.[16, 24, 25] In contrast, my definition (ii) is unambiguous in such situations. It yields well-defined statements when we are given results of measurements both in the past and in the future of P_i and in cases when the space–time points P_i are such that future and past cannot be unambiguously defined.

For a simple time-symmetric case in which $\mathscr{M}'$ describes a single measurement of a variable A performed between two complete measurements which fix the states $|\Psi_1\rangle$ at t_1 and $|\Psi_2\rangle$ at t_2, definition (ii) becomes the following.

> (iii) *If a measurement of an observable A has been performed at time t, $t_1 < t < t_2$, then the probability for $A = a_i$ would be equal to p_i, provided that the results of measurements performed on the system at times t_1 and t_2 are fixed.*

The probabilities p_i are given by the ABL formula:[26, 27]

$$\mathrm{Prob}(a_i) \equiv p_i = \frac{|\langle \Psi_2 | \mathbf{P}_{A=a_i} | \Psi_1 \rangle|^2}{\sum_j |\langle \Psi_2 | \mathbf{P}_{A=a_j} | \Psi_1 \rangle|^2} \tag{1}$$

The application of the time-symmetric formula (1) to counterfactual situations led to considerable controversy.[28–35] I believe that the time-symmetric definition (iii) provides a consistent way for application of the ABL rule for counterfactual situations, thus resolving the controversy.

3. ELEMENTS OF REALITY

Definition (iii) is also helpful in analyzing various attempts to prove that *realistic* quantum theory leads to a contradiction with relativistic

causality. The "element of reality" can be considered as an example of a counterfactual (iii) in the particular case of probability 1 for a certain outcome. A time-symmetrized definition of element of reality is as follows.[36]

> (iv) *If we can infer with certainty that the result of a measurement at time t of an observable A is a, then, at time t, there exists an element of reality A = a.*

The word "infer" is neutral relative to past and future. The inference about results at time t is based on the results of measurements on the system performed both before and after time t.

An important feature of time-symmetric elements of reality (iv) of a pre- and postselected quantum system is that the "product rule" does not hold. The product rule means that if $A = a$ and $B = b$ are elements of reality, then $AB = ab$ is also an element of reality.

A simple example of this hind is a system of two spin-$\frac{1}{2}$ particles prepared at t_1 in a singlet state,

$$|\Psi_1\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle_1 |\downarrow\rangle_2 - |\downarrow\rangle_1 |\uparrow\rangle_2) \tag{2}$$

At t_2 the particles are found in the state

$$|\Psi_2\rangle = |\uparrow_x\rangle_1 |\uparrow_y\rangle_2 \tag{3}$$

A set of elements of reality for these particles at an intermediate time t is [use the ABL formula (1) to see this]

$$\{\sigma_{1y}\} = -1 \tag{4}$$

$$\{\sigma_{2x}\} = -1 \tag{5}$$

$$\{\sigma_{1y}\sigma_{2x}\} = -1 \tag{6}$$

where the notation $\{X\}$ signifies the outcome of a measurement of X. Indeed, the product rule does not hold: $\{\sigma_{1y}\sigma_{2x}\} \neq \{\sigma_{1y}\}\{\sigma_{2x}\}$. Note that a measurement of the nonlocal variable in Eq. (6), the product of local variables related to separated locations, is not disallowed due to locality of physical interactions. This particular measurement can be performed using local interactions only.[37]

The failure of the product rule plays an important role in discussing Lorentz invariance of a realistic quantum theory, especially, in the light of recent proposals to prove the impossibility of a realistic Lorentz invariant quantum theory that applied the product rule,[24, 25] which generated considerable controversy.[38–41]

4. ANALYSIS OF STAPP'S NONLOCALITY ARGUMENT

It seems to me that the proposed definition for counterfactuals should also help to resolve the recent controversy generated by the proposal of Stapp[16] mentioned above. My definition resolves the vagueness in these discussions, pointed out by Finkelstein,[22] about what is fixed in the counterfactual worlds.

I claim that quantum theory does not support the second locality condition of Stapp[16] (his LOC2). Stapp considers two spatially separated spin-$\frac{1}{2}$ particles. In his example, a certain counterfactual statement related to a particle on the right can be proved given that a certain action was performed before that on a spatially separated particle on the left. He then notes that in another Lorentz frame the action on the particle on the left is performed *after* the time to which the counterfactual statement is related. Stapp concludes that since an action in the future cannot influence the past, the action on the left side can be replaced by some other action without changing the truth of the counterfactual related to the particle on the right.

The argument which led Stapp to his locality condition LOC2 does not go through if we adopt the definition of counterfactuals (ii), considering measurements on the particle on the right while keeping fixed the results of all other measurements on our system (the system consisting ot the two spin-$\frac{1}{2}$ particles). Then the truth of the counterfactual requires only the existence of a Lorentz frame in which the measurements on the right side are after the measurement on the left, and the consideration of the other Lorentz frames is irrelevant.

In Stapp's example we indeed have a situation in which an action in a space-like separated region on the left side changes the truth of a certain counterfactual statement about measurement on the right. However, I do not see that the failure of LOC2 proves the "nonlocal character of quantum theory" as the title and the spirit of Stapp's paper suggest. In order to demonstrate the meaning of LOC2, let me present another example where it fails.

Consider again two spatially separated spin-$\frac{1}{2}$ particles prepared in a singlet state (2). At time t a Stern–Gerlach experiment with the gradient of a magnetic field in the positive $\hat{z}$ direction is performed and the result $\sigma_{2z} = \alpha$ is obtained. Now consider the following counterfactual statement.

> **CF**: *If the measurement were performed with the gradient pointing in the negative* $\hat{z}$ *direction instead, the same result,* $\sigma_{2z} = \alpha$, *would be obtained.*

The truth of this statement depends on actions on particle 1 in a space-like separated region: if the measurement of σ_{1z} were performed, then CF would be true; if no measurement were performed or, say, σ_{1x} were measured instead, then the truth of CF would not follow. Indeed, if σ_{1z} was measured, then, in the actual world (in which $\sigma_{2z} = \alpha$ was obtained), the outcome would be $\sigma_{1z} = -\alpha$. Since in a counterfactual world the results of measurements in space-like regions are not changed, the outcome of $\sigma_{2z} = \alpha$ must be found in all counterfactual worlds, irrespective of the type of measuring device which is used for this measurement. In our case, a counterfactual world with a reversed gradient of the magnetic field, the outcome $\sigma_{2z} = \alpha$ corresponds to the spot in the opposite location. If σ_{1x} were measured instead, then irrespective of the outcome of this measurement, both results of the measurement of σ_{2z}, α and $-\alpha$, would be possible, and therefore, CF might not be true.

Note the even more dramatic difference in the framework of the Bohm–Bell hidden-variable interpretation:[42, 43] CF is true if σ_{1z} is measured and CF is false if σ_{1x} is measured. Indeed, if σ_{1x} is measured, the assumption of the same initial hidden variables (the Bohmian positions) in the actual and the counterfactual worlds leads to the spot in the same position in the Stern–Gerlach experiment even in the case of a reversed gradient of the magnetic field. The same spot with a reversed gradient corresponds to the opposite result of the measurement of σ_{2z}. Therefore, CF is false. In the case of σ_{1z} measurement, the standard quantum theory yields definite prediction about the σ_{2z} measurement: CF is true. Therefore, all valid interpretations, including Bohmian interpretation, must yield the same prediction.

Of course, since CF cannot be tested, no contradiction with relativistic causality can arise. Still, there is some nonlocality in this example. For me, the framework of the MWI yields the clearest picture of this nonlocality. By performing measurements on particle 1, we split our world into two worlds, creating a *mixture* of two worlds for particle 2. With different choices of measurement on particle 1, we create different mixtures of worlds for particle 2, for example, two worlds with definite σ_{2z} (for which CF is true) or two worlds with definite σ_{2x} (for which CF does not follow). Although the worlds are different, the two mixtures are physically equivalent for particle 2 and therefore there was no nonlocal action in the physical universe which incorporates all the worlds. The nonlocality is as follows: the world (branch) in the MWI is a nonlocal entity which, in our case, is defined by properties in the location of the two particles. The choice of a local measurement on particle 1 defines the set of worlds into which the present world will be split. In this way an action on particle 1 leads to various sets of possible properties related to particle 2.

5. APPARENT WEAKNESSES OF THE TIME-SYMMETRIZED QUANTUM COUNTERFACTUALS

I have to mention a property of definition (ii) which might be considered its weakness. The outcome of measurement $\mathcal{M}$ performed in the actual world plays no role in calculating the truth of the counterfactual statement (except trivial cases in which $\mathcal{P}$ involves a comparison between the outcome of $\mathcal{M}$ and that of $\mathcal{M}'$ as in the previous example). It is assumed that properties of the outcome of $\mathcal{M}'$ are independent on the outcome of $\mathcal{M}$. This is what standard quantum theory tells us, but this is not true, in general, for hidden-variable theories: the outcome of $\mathcal{M}$ can yield certain information about hidden variables, information which might help to ascertain the properties of the outcome of $\mathcal{M}'$. In the framework of the hidden-variables theories, definition (ii) is incomplete; we must add a statement about hidden variables, for example, by fixing hidden variables in a counterfactual world to be equal to the hidden variables in the actual world. I have adopted this approach two paragraphs above, but it should be noted that it is explicitly time-asymmetric: the hidden variables are fixed only in the past. I do not know how to approach the problem of time-symmetric hidden variables.

The proposed definitions of counterfactuals (ii) and (iii) are also applicable for counterfactuals in classical physics. However, due to the determinism of classical theory, we cannot fix independently the results of a complete set of measurements in the past and the results of the complete set of measurements in the future. Note that there are certain limitations of this kind in the quantum case too. For example, consider a spin-$\frac{1}{2}$ particle with three consecutive measurements, $\sigma_z(t_1) = 1$, $\sigma_x(t) = 1$, and $\sigma_z(t_2) = -1$, $t_1 < t < t_2$. Then a counterfactual statement, "If at time t a measurement of σ_z were performed instead, the result would be $\sigma_z(t) = 1$," is neither true or false, but meaningless, because the results of measurements $\sigma_z(t_1) = 1$ and $\sigma_z(t_2) = -1$ are impossible when σ_z, instead of σ_x, is measured at time t. Nevertheless, such constrains are not strong and they leave room for numerous nontrivial counterfactuals.

In classical physics the counterfactuals (ii) have an even more serious problem. $\mathcal{M}'$ consists of measurements of some observables. We can make a one-to-one correspondence between "The outcome of a measurement of an observable O is o_i" and "The value of O is o_i." The latter is independent of whether or not the measurement of O has been performed, and therefore, statements which are formally counterfactual about results of possible measurements can be replaced by "factual" (unconditional) statements about values of corresponding observables. In contrast, in standard quantum theory, observables, in general, do not have definite values and therefore

we cannot always reduce the above counterfactual statements to "factual" statements.

I do not expect that everybody will agree with my proposals for resolving the controversies discussed above. I hope only that the main result of this work will not be controversial: a consistent definition of counterfactuals in quantum theory, a definition that is equivalent to the standard approach for the time-asymmetric cases in which only the past of the system is given, but that is applicable to the time-symmetric situation (such as pre- and postselected systems)—a definition which is a useful tool for the analysis of many current problems.

ACKNOWLEDGMENTS

This research was supported fin part by Grants 614/95 and 471/98 from the Basic Research Foundation (administered by the Israel Academy of Sciences and Humanities).

REFERENCES

1. L. Vaidman, e-print quant-ph/9802042 (1998).
2. R. E. Kastner, *Found. Phys.* **29** (6) (1999), to appear.
3. D. Lewis, *Nous* **13**, 455 (1979); reprinted in D. Lewis, *Philosophical Papers*, *Vol. II* (Oxford University Press, Oxford, p. 32).
4. J. Bennett, *Phil. Rev.* **93**, 57 (1984).
5. R. Penrose, *Shadows of the Mind* (Oxford University Press, Oxford, 1994).
6. A. Elitzur and L. Vaidman, *Found. Phys.* **23**, 987 (1993).
7. H. Everett, *Rev. Mod. Phys.* **29**, 454 (1957).
8. L. Vaidman, *Int. Stud. Phil. Sci.* **12**, 245 (1998).
9. D. Deutsch, *The Fabric of Reality* (Penguin, New York, 1997).
10. L. Vaidman, *Phil. Sci. Assoc.* **1994**, 211 (1994).
11. B. Skyrms, *Phil. Sci.* **49**, 43 (1982).
12. A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer Academic, Dordrecht, 1993).
13. N. D. Mermin, in *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem*, J. T. Cushing and E. McMullin, eds. (University of Notre Dame Press, Notre Dame, IN, 1989).
14. D. Bedford and H. P. Stapp, *Synthese* **102**, 139 (1995).
15. D. Lewis, *Counterfactuals* (Blackwell, Oxford, 1973).
16. H. P. Stapp, *Am. J. Phys.* **65**, 300 (1997).
17. L. Hardy, *Phys. Lett. A* **167**, 17 (1992).
18. N. D. Mermin, *Am. J. Phys.* **66**, 920 (1998).
19. H. P. Stapp, *Am. J. Phys.* **66**, 924 (1998).
20. W. Unruh, *Phys. Rev. A* **59**, 126 (1999).
21. H. P. Stapp, e-print quant-ph/9801056 (1998).
22. J. Finkelstein, e-print quant-ph/9801011 (1998).

23. V. S. Mashkevich, e-print quant-ph/9801032 (1998).
24. L. Hardy, *Phys. Rev. Lett.* **68**, 2981 (1992).
25. R. Clifton, C. Pagonis, and I. Pitowsky, *Phil. Sci. Assoc.* 1992 **1**, 114 (1994).
26. Y. Aharonov, P. G. Bergmann, and J. L. Lebowitz, *Phys. Rev. B* **134**, 1410 (1964).
27. Y. Aharonov and L. Vaidman, *J. Phys. A* **24**, 2315 (1991).
28. D. Albert, Y. Aharonov, and S. D'Amato, *Phys. Rev. Lett.* **54**, 5 (1985).
29. J. Bub and H. Brown, *Phys. Rev. Lett.* **56**, 2337 (1986).
30. W. D. Sharp and N. Shanks, *Phil. Sci.* **60**, 488 (1993).
31. O. Cohen, *Phys. Rev. A* **51**, 4373 (1995).
32. L. Vaidman, *Phys. Rev. A* **57**, 2251 (1998).
33. D. J. Miller, *Phys. Lett. A* **222**, 31 (1996).
34. R. E. Kastner, *Stud. Hist. Phil. Mod. Phys.* **30**, 237 (1999).
35. L. Vaidman, *Stud. Hist. Phil. Mod. Phys.*, in press (e-print quant-ph/9811092, Sept. 1999).
36. L. Vaidman, *Phys. Rev. Lett.* **70**, 3369 (1993).
37. Y. Aharonov, D. Albert, and L. Vaidman, *Phys. Rev. D* **34**, 1805 (1986).
38. L. Vaidman, in *Symposium on the Foundations of Modern Physics*, P. J. Lahti, P. Bush, and P. Mittelstaedt, eds. (World Scientific, Singapore, 1993), p. 406.
39. L. Vaidman, e-print quant-ph/9703018 (1997).
40. O. Cohen and B. J. Hiley, *Phys. Rev. A* **52**, 76 (1995).
41. O. Cohen and B. J. Hiley, *Found. Phys.* **26**, 1 (1996).
42. D. Bohm, *Phys. Rev.* **85**, 97 (1952).
43. J. S. Bell, *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge, 1987), p. 117.