Pergamon

Stud. Hist. Phil. Mod. Phys., Vol. 30, No. 3, pp. 373–397, 1999

Printed in Great Britain
1355-2198/99 $ - see front matter

Defending Time-Symmetrised Quantum Counterfactuals

*Lev Vaidman**

Recently, several authors have criticised the time-symmetrised quantum theory originated by the work of Aharonov, Bergmann and Lebowitz (1964). The core of this criticism was a proof, appearing in various forms, which showed that the counterfactual interpretation of time-symmetrised quantum theory cannot be reconciled with standard quantum theory. I (Vaidman, 1996a, 1997) have argued that the apparent contradiction is due to a logical error and have introduced consistent time-symmetrised quantum counterfactuals. Here I repeat my arguments defending the time-symmetrised quantum theory and reply to the criticism of these arguments by Kastner (1999). © 1999 Published by Elsevier Science Ltd. All rights reserved.

1. Introduction

Starting from the seminal work of Aharonov, Bergman and Lebowitz (ABL) (1964), Aharonov, myself and others are developing a time-symmetrised formalism of quantum theory (TSQT). Recently a particular question related to this formalism, namely the validity of the counterfactual application of the ABL rule, became a subject of a significant controversy culminating in the paper by Kastner (1999). According to the critics some of the recent results obtained in the framework of the TSQT are based on the counterfactual interpretation of the ABL rule which, in general, is inconsistent. In recent papers (Vaidman, 1996a, 1997) I have defended the TSQT and, moreover, have introduced time-symmetrised counterfactuals for quantum theory for which the ABL rule is valid. Kastner critically analyses my papers and claims that my defense is not sound. In this paper I refute Kastner's arguments. For completeness I include the relevant sections of the papers.

(*Received 22 November 1998; revised 22 February 1999*)
* School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel (*e-mail*: vaidman@post.tau.ac.il).

PII: S1355-2198(99)00013-1

Lewis, who is probably the main authority on counterfactuals, writes: 'Counterfactuals are infected with vagueness, as everybody agrees' (1986, p. 34). I do not completely agree. I believe that quantum counterfactuals can be defined unambiguously (as I will do in Section 3). However, it seems that the core of the current controversy is indeed the ambiguity of the concept of counterfactuals. Kastner distinguishes between two apparently counterfactual readings of the ABL rule: the first one she considers as 'non-counterfactual' and the second is a '*bona fide* counterfactual' (1999, p. 239). She claims that the second reading is what is frequently applied in the TSQT. I will argue below that it is the first reading which is correct and which has been applied in the framework of the TSQT. The question whether the first reading is 'counterfactual' remains a semantic issue. Formally, it is. Moreover, I will argue below that Kastner's second reading is inconsistent and that all other proposed counterfactual readings of the ABL rule are either inconsistent or not time-symmetric. Thus, I find it appropriate to use the term 'counterfactual' for my (Kastner's first) reading. However, if philosophers find it important to spell out the differences between these 'non-counterfactual' counterfactuals and '*bona fide* counterfactuals' applied in general analyses, I hope that the current discussion will help in this task.

The plan of this paper is as follows. In Section 2 I briefly define the time-symmetrised formalism.[1] In section 3 I analyse the concept of counterfactuals in quantum theory and introduce the time-symmetrised counterfactuals. In Section 4 I discuss *elements of reality* which are examples of quantum counterfactuals. Section 5 is devoted to the analysis of the inconsistency proof of Sharp and Shanks (1993) and its variations. Section 6 presents a more detailed analysis of possible counterfactual interpretations of the ABL rule. In Section 7 I analyse Kastner's readings of the ABL rule, in Section 8 her analysis of the Sharp and Shanks proof, and in Section 9 her criticism of my definition of time-symmetrised counterfactuals. Section 10 summarises the arguments of the paper.

2. Time-Symmetrised Formalism

In standard quantum theory a complete description of a system at a given time is given by a quantum state $|\Psi\rangle$. It yields the probabilities for all results c_j of a measurement at that time of any observable C according to the equation

$$P(c_j) = |\langle\Psi|\mathbf{P}_{C=c_j}|\Psi\rangle|, \tag{1}$$

where $\mathbf{P}_{C=c_j}$ is the projection operator on the subspace defined by $C = c_j$. Although it is not manifestly apparent, equation (1) is intrinsically asymmetric in time: the state $|\Psi\rangle$ is determined by some measurements in the past and it evolves toward the future. The time evolution between the measurements, however, is considered time-symmetric since it is governed by the Schrödinger

[1] Sections 2–5 are from Vaidman (1997) and Section 6 from Vaidman (1996a). These are the papers which Kastner criticises. The text is slightly revised: notation is unified and some English corrections are made. The full original text of the papers is available electronically.

equation, for which each solution evolving forward in time has its counterpart (its complex conjugate with some other well understood simple changes) evolving backward in time. The asymmetry in time of the standard quantum formalism is manifested in the absence of the quantum state evolving backward in time from future measurements (relative to the time in question).

Time-symmetrised quantum theory completely describes a system at a given time by a two-statevector $\langle\Psi_2||\Psi_1\rangle$. It yields the (conditional) probabilities for all results c_j of a measurement of any observable C at that time according to the generalisation of the ABL formula (Aharonov and Vaidman, 1991):

$$P(c_j) = \frac{|\langle\Psi_2|\mathbf{P}_{C=c_j}|\Psi_1\rangle|^2}{\sum_i |\langle\Psi_2|\mathbf{P}_{C=c_i}|\Psi_1\rangle|^2}. \tag{2}$$

The time symmetry means that $\langle\Psi_2|$ and $|\Psi_1\rangle$ enter the equations, and thus govern the observable results, on an equal footing. Moreover, the time symmetry means that, in regard to time-symmetric measurements, a system described by the two-statevector $\langle\Psi_2||\Psi_1\rangle$ is identical to a system described by the two-statevector $\langle\Psi_1||\Psi_2\rangle$. I analyse the time symmetry of the process of measurement in Section 6 of (Vaidman, 1997); here I only point out that ideal measurements are time-symmetric. Indeed, the symmetry under the interchange of $\langle\Psi_2|$ and $|\Psi_1\rangle$ is explicit in equation (2) which refers to ideal measurements.

Another basic concept of the time-symmetrised two-statevector formalism is *weak value*. An (almost) standard measurement procedure for measuring observable C with weakened coupling (which we (Aharonov and Vaidman, 1990) call *weak measurement*) yields the *weak value* of C:

$$C_w \equiv \frac{\langle\Psi_2|C|\Psi_1\rangle}{\langle\Psi_2|\Psi_1\rangle}. \tag{3}$$

Here again, $\langle\Psi_2|$ and $|\Psi_1\rangle$ enter the equations on an equal footing. However, when we interchange $\langle\Psi_2|$ and $|\Psi_1\rangle$, the weak value changes to its complex conjugate. Thus, in this situation, as for the Schrödinger equation, time reversal is accompanied by complex conjugation.

In order to explain how to obtain a quantum system described at a given time t by a two-statevector $\langle\Psi_2||\Psi_1\rangle$, we shall assume for simplicity that the free Hamiltonian of the system is zero. In this case, it is enough to prepare the system at time t_1 prior to time t in the state $|\Psi_1\rangle$, and to ensure no disturbance between t_1 and t as well as between t and t_2, and to find the system at t_2 in the state $|\Psi_2\rangle$. It is crucial that $t_1 < t < t_2$, but the relation between these times and 'now' is not fixed. The times t_1, t, t_2 might all be in the past, or we can discuss future measurements and then they are all in the future; we just have to agree to discard all cases when the measurements at time t_2 does not yield the result corresponding to the state $|\Psi_2\rangle$.

Note the asymmetry between the measurement at t_1 and the measurement at t_2. Given an ensemble of quantum systems, it is always possible to prepare all of them in a particular state $|\Psi_1\rangle$, but we cannot ensure finding the system in

a particular state $|\Psi_2\rangle$. Indeed, if the pre-selection measurement yielded a result different from projection on $|\Psi_1\rangle$ we can always change the state to $|\Psi_1\rangle$, but if the measurement at t_2 did not show $|\Psi_2\rangle$, our only choice is to discard such a system from the ensemble. This asymmetry, however, is not relevant to the problem we consider here. We study the symmetry relative to the measurements at time t for a *given* pre- and post-selected system, and we do not investigate the time symmetry of obtaining such a system. The only important detail is that the interaction at time t has to be time-symmetric. See Section 6 of Vaidman (1997) for a more detailed discussion of these issues.

3. Counterfactuals

A general form of a counterfactual statement is

Definition (i). If it were that A, then it would be that B.

There are many philosophical discussions of the concept of counterfactuals and especially on time's arrow in counterfactuals. Many of the discussions (e.g. Lewis, 1986; Bennett, 1984) are related to A: how come A if in the actual world A is not true? Do we need a miracle (a violation of a fundamental law of nature) for A? Does A come by itself, or it is accompanied by other changes? However, these questions are not relevant to the problem of counterfactuals in quantum theory. The questions about A are not relevant because A depends solely on an external system, which is not under discussion by the definition of the problem. Indeed, in quantum theory the counterfactuals have a very specific form:[2]

$$A = \text{a measurement } M \text{ is performed,}$$

$$B = \text{the result of } M \text{ has property } P.$$

The measurement M might consist of measurements of several observables performed together. The property P might be a certain relation between the results of measurements of these observables or a probability for a certain relation or for a certain result.

It is assumed that the experimenter can make any decision about which measurement to perform and the question how he makes this decision is not considered. It is assumed that the experimenter and his measuring devices are not correlated in any way with the state of the system prior to the measurement. Thus, in the world of the quantum system no miracles are needed and no changes relative to the actual world have to be made for different A's.[3]

[2] This definition of counterfactuals in quantum theory is broad enough for discussing issues relevant to this paper. However, in some cases the term 'counterfactuals' has been used differently. For example, in Penrose (1994, p. 240) '*counterfactuals* are things that might have happened, although they did not in fact happen'.

[3] The indeterminism of standard quantum theory allows us even to discuss worlds which include the experimenter without invoking miracles. As an example, consider an experimenter who chooses between different measurements according to a random result of another quantum experiment.

Although one can define counterfactuals of this form in the framework of classical theory, they are of no interest because they are equivalent to some 'factual' statements. In classical physics any observable always has a definite value and a measurement of the observable yields this value. Therefore, there is a one-to-one correspondence between 'the result of a measurement of an observable C is c_j' and 'the value of C is c_j'. The latter is independent of whether the measurement of C has been performed or not 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 do not in general have definite values and therefore we cannot always reduce the above counterfactual statements to 'factual' statements.

Most of the discussions of counterfactuals in quantum theory are in the context of EPR–Bell-type experiments. Some examples are Skyrms (1982), Peres (1993), Mermin (1989) (which, however, does not use the word 'counterfactual'), Ghirardi and Grassi (1994) and Bedford and Stapp (1995): the last of these presents an analysis of a Bell-type argument in the formal language of Lewis' (1973) theory of counterfactuals. In these discussions, the common scenario is that a composite system is described at a certain time by some entangled state and then an array of incompatible measurements on this system at a later time is considered. Various conclusions are derived from statements about the results of these measurements. Since these measurements are incompatible they cannot all be performed together, so that necessarily at least some of them were not actually performed. This is why they are called counterfactual statements.

These counterfactuals are explicitly asymmetric in time. The asymmetry is neither in A nor in B; both are about a single time t. The asymmetry is in the description of the actual world. The *past* and not the *future* (relative to t) of a system is given.

This, however, is not the only asymmetry of the counterfactuals in quantum theory as they are usually considered. A different asymmetry (although it looks very similar) is in what we assume to be 'fixed', i.e. which properties of the actual world we assume to be true in possible counterfactual worlds. The *past* and not the *future* of the system is fixed.

It seems that while the first asymmetry can be easily removed, the second asymmetry is unavoidable. According to standard quantum theory a system is described by its quantum state. In the actual world, in which a certain measurement has been performed at time t (or no measurement has been performed at t) the system is described by a certain state before t, and by some state after time t. In the counterfactual world in which a different measurement was performed at time t, the state before t is, of course, the same, but the state after time t is invariably different (if the observables measured in actual and counterfactual worlds have different eigenstates). Therefore, we cannot hold fixed the quantum state of the system in the future.[4]

[4] Note that none of these asymmetries exists in the classical case because when a complete description of a classical system is given at one time, it fixes the complete description at all times and (ideal) measurements at time t do not change the state of a classical system.

The argument above shows that for constructing time-symmetric counterfactuals we have to give up the description of a quantum system by its quantum state. Fortunately we can do that without loosing anything except the change due to the measurement at time t which caused the difficulty. A quantum state at a given time is completely defined by the results of a complete set of measurements performed prior to this time. Therefore, we can take the set of all results performed on a quantum system as a description of the world of the system instead of describing the system by its quantum state. (This proposal will also help to avoid ambiguity and some controversies related to the description of a single quantum system by its quantum state.) Thus, I propose the following definition of counterfactuals in the framework of quantum theory:

Definition (ii). If a measurement M were performed at time t, then it would have property P, provided that the results of all measurements performed on the system at all times except the time t are fixed.

For time-asymmetric situations in which only the results of measurements performed before t are given (and thus only these results are fixed), this definition of counterfactuals is equivalent to the counterfactuals as they have usually been used. However, when the results of measurements performed on the system both before and after the time t are given, Definition (ii) yields novel time-symmetrised counterfactuals. In particular, for the ABL case, in which *complete* measurements are performed on the system at t_1 and t_2, $t_1 < t < t_2$, we obtain:

Definition (iii). If a measurement of an observable C were performed at time t, then the probability for $C = c_j$ would equal $P(c_j)$, provided that the results of measurements performed on the system at times t_1 and t_2 are fixed.

The ABL formula (2) yields correct probabilities for counterfactuals defined as in Definition (iii). That is: in the experiment in which C is measured at time t on the systems from a pre- and post-selected ensemble defined by fixed results of the measurements at t_1 and t_2 (all such systems and only such systems are considered), the frequency of result c_j is $P(c_j)$, where $P(c_j)$ is as given by (2).

For the ABL situation one can also define a time-*asymmetric* counterfactual:

Definition (iv). Given the results of measurements at t_1 and t_2, $t_1 < t < t_2$ (in the actual world), if a measurement of an observable C were performed at time t, then the probability for $C = c_j$ would equal $P(c_j)$, provided that the results of all measurements performed on the system at all times before time t are fixed.

In the framework of standard quantum theory the information about the result of measurement at t_2 is irrelevant: the probability for $C = c_j$ does not depend on this result. Thus, it is obvious that the ABL formula (2), which includes the result of the measurement at time t_2 explicitly, does not yield counterfactual probabilities according to Definition (iv).

One might modify Definition (iv) in the framework of some 'hidden variable' theory with a natural additional requirement of fixing the hidden variables of the system in the past. The properties of such counterfactuals will depend

crucially on the details of the hidden variable theory (see the discussion by Aharonov and Albert (1987) in the framework of Bohm's theory), but the ABL formula (2) is not valid for any such modification. In order to show this consider a spin-$\frac{1}{2}$ particle which was found at t_1 and at t_2 in the same state $|\uparrow_z\rangle$ (and no measurement has been performed at t). We ask what is the (counterfactual) probability for finding spin 'up' in the direction $\hat{\xi}$ which makes an angle θ with the direction $\hat{z}$, at the intermediate time t. In this case, hidden variables, even if they exist, cannot change that probability because any particle found at t_1 in the state $|\uparrow_z\rangle$, irrespectively of its hidden variable, yields the result 'up' in the measurement at t_2. Therefore, the statistical predictions about the intermediate measurement at time t must be the same as for the pre-selected-only ensemble (these are *identical* ensembles in this case), i.e.

$$P(\uparrow_\xi) = |\langle\uparrow_\xi|\uparrow_z\rangle|^2 = \cos^2(\theta/2). \tag{4}$$

The ABL formula, however, yields:

$$P(\uparrow_\xi) = \frac{|\langle\uparrow_z|\mathbf{P}_{\uparrow_\xi}|\uparrow_z\rangle|^2}{|\langle\uparrow_z|\mathbf{P}_{\uparrow_\xi}|\uparrow_z\rangle|^2 + |\langle\uparrow_z|\mathbf{P}_{\downarrow_\xi}|\uparrow_z\rangle|^2} = \frac{\cos^4(\theta/2)}{\cos^4(\theta/2) + \sin^4(\theta/2)}. \tag{5}$$

The fact that the ABL formula (2) does not hold for the counterfactuals defined in (iv) or its modifications is not surprising. Definition (iv) is explicitly asymmetric in time. The ABL formula, however, is time-symmetric and therefore it can hold only for time-symmetrised counterfactuals.

A recent study of time's arrow and counterfactuals in the framework of quantum theory by Price (1996) seems to support my Definition (ii). Let me quote from his section 'Counterfactuals: What Should We Fix?':

> Hold fixed the past, and the same difficulties arise all over again. Hold fixed merely what is accessible, on the other hand, and it will be difficult to see why this course was not chosen from the beginning (1996, p. 179).

This quotation looks very much like my proposal. Indeed, I find many arguments in his book pointing in the same direction. However, in fact, this quotation represents a time asymmetry: according to Price 'merely what is accessible' is 'an accessible past'. But this is not the time asymmetry of the physical theory; Price writes that 'no physical asymmetry is required to explain it'. Although the book includes an extensive analysis of a photon passing through two polarisers — the classic set-up for the ABL formula — I found no explicit discussion of a possible measurement in between the polarisers, i.e. of the problem we discuss here.[5]

[5] Price briefly and critically mentions the ABL paper. He writes (1996, p. 208): 'What they [ABL] fail to note, however, is that their argument does nothing to address the problem for those who disagree with Einstein — those who think that the state function is a complete description, so that the change that takes place on measurements is a real change in the world, rather than merely a change in our knowledge of the world'. This seems to me an unfair criticism: ABL clearly state that in the situations they consider 'the complete description' is given by *two* wave functions (see Aharonov and Vaidman (1991) for a more detailed discussion). Moreover, it seems to me that the development of this time-symmetrised quantum formalism is not too far from the spirit of 'advanced action' — Price's vision of the solution of the problem of time's arrow.

4. Elements of Reality

Among the important counterfactuals in quantum theory are 'elements of reality'. For comparison, I will start with a definition of a time-asymmetric element of reality:

> *Definition* (v). If we can *predict* with certainty that the result of measuring at time t an observable C is c, then, at time t, there exists an element of reality $C = c$.

This is, essentially, a quotation from Redhead (1987, p. 72). However, there is a significant difference: I consider this counterfactual sentence as the *definition* of the concept of elements of reality, while Redhead considers it as a sufficient condition for the existence of an element of reality. Redhead was inspired by the criterion for elements of reality by Einstein, Podolsky and Rosen (EPR). In spite of the similarity in its form, the EPR criterion is very different: 'If, without in any way disturbing the system, we can predict with certainty the value of a physical quantity [...] '. The crucial difference is that 'predict' in the EPR criterion means to find out using certain (non-disturbing) measurements, while in my definition 'predict' means to deduce using existing information. Thus, for two spin-$\frac{1}{2}$ particles in a singlet state, the value of a spin component of a single particle in any direction is an element of reality in the EPR sense (it can be found out by measuring another particle), while there is no element of reality for a spin component value in any direction according to my definition (in the EPR state, the probability to find spin 'up' in any direction is $\frac{1}{2}$).

Definition (v) of elements of reality is asymmetric in time because of the word 'predict'. I have proposed a modification of this definition applicable for time-symmetric elements of reality (Vaidman, 1993):

> *Definition* (vi). If we can *infer* with certainty that the result of measuring at time t an observable C is c, then, at time t, there exists an element of reality $C = c$.

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. Note that in some situations we can 'infer' more facts than can be obtained by 'prediction' based on the results in the past and 'retrodiction' based on the results in the future (relative to t) together.[6]

The difference between definitions of 'elements of reality', (v) and (vi), and definitions of counterfactuals in quantum theory, (iii) and (iv), is that the

[6] For example, for a spin-$\frac{1}{2}$ particle, results of measurements in the past can lead to prediction of a certain result of a spin-component measurement at most for a single direction. The same is true for retrodiction based on the results of measurements in the future and, therefore, prediction and retrodiction can lead to certain results of spin-component measurements at the intermediate time for, at most, two directions. Nevertheless, in the example given by Vaidman *et al.* (1987, p. 15), inference from particular results of measurements in the past and in the future taken together lead to certain results of spin component measurements for a continuum of directions.

property P in (v) and (vi) is constrained to be 'the result of measuring at time t of an observable C is c'. In fact, time-asymmetric 'elements of reality' (v), defined as 'predictions', do not represent 'interesting' counterfactuals. There is no non-trivial set of such counterfactual statements, i.e. any set of such statements can be tested on a single system. Indeed, all observables the measurement of which yield some results with certainty for a pre-selected system can be tested together. One way to extend the definition of time-asymmetric elements of reality in order to get non-trivial counterfactuals is to consider 'multiple-time measurements' (instead of measurements at time t only). Another extension, which corresponds to numerous analyses in the literature (e.g. references on quantum counterfactuals mentioned above), is to go beyond statements about observables which have definite values:

Definition (vii). If we can *predict* with certainty a relation between the results c_α of measuring at time t a set of observables C_α, then, at time t, there exists a 'generalised element of reality' which is this relation between the c_α's.

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

$$|\Psi_1\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle_1|\downarrow\rangle_2 - |\downarrow\rangle_1|\uparrow\rangle_2). \tag{6}$$

We can predict with certainty that the results of measurements of spin components of the two particles fulfill the following two relations:

$$\{\sigma_{1x}\} + \{\sigma_{2x}\} = 0, \tag{7}$$

$$\{\sigma_{1y}\} + \{\sigma_{2y}\} = 0, \tag{8}$$

where $\{\sigma_{1x}\}$ signifies the result of a measurement of the x-component of the spin of the first particle, etc. The relations (7) and (8) represent a set of generalised elements of reality of type (vii). This is a non-trivial set of counterfactuals because (7) and (8) cannot be tested together: the measurement of σ_{1x} disturbs the measurement of σ_{1y} just as the measurement of σ_{2x} disturbs the measurement of σ_{2y}.

In contrast, the set of elements of reality of type (v) given by

$$\{\sigma_{1x} + \sigma_{2x}\} = 0, \tag{9}$$

$$\{\sigma_{1y} + \sigma_{2y}\} = 0 \tag{10}$$

can be tested on a single system; see Aharonov *et al.* (1986) for a description of such measurements. Yet another set of counterfactuals, which consists of definite statements about measurements, but which does not belong to the type (v) because they concern *two-time* measurements performed at two different times t_1 and t_2, *cannot* be tested on a single system:

$$\{\sigma_{1x}(t_1) + \sigma_{2x}(t_2)\} = 0, \tag{11}$$

$$\{\sigma_{1y}(t_2) + \sigma_{2y}(t_1)\} = 0. \tag{12}$$

Note also a situation which involves only a single free spin-$\frac{1}{2}$ particle. The particle is prepared, before t_1, $t_1 < t_2 < t_3$, in the state $|\uparrow_y\rangle$. Then, a nontrivial set of counterfactuals is:

$$\{\sigma_x(t_1) - \sigma_x(t_3)\} = 0, \tag{13}$$

$$\{\sigma_y(t_2)\} = 1. \tag{14}$$

In this example, however, statement (13) has a somewhat different character because it depends not on the results of measurements performed on the particle before or after the period of time (t_1, t_2), but on the fact that the system was not disturbed during this period of time.

5. Inconsistency Proofs

The key point of the criticism of the time-symmetrised quantum theory (Sharp and Shanks, 1993; Cohen, 1995; Miller, 1996) is the conflict between counterfactual interpretations of the ABL rule and predictions of quantum theory. I shall argue here that the inconsistency proofs are unfounded and therefore the criticism essentially falls apart.

The structure of all these inconsistency proofs is as follows. Three consecutive measurements are considered. The first is the preparation of the state $|\Psi_1\rangle$ at time t_1. The probabilities for the results c_j of the second measurement at time t are considered. The final measurement at time t_2 is introduced in order to allow the analysis using the ABL formula. Sharp and Shanks consider three consecutive spin-component measurements of a spin-$\frac{1}{2}$ particle in different directions. Cohen analyses a particular single-particle interference experiment. It is a variation of the Mach–Zehnder interferometer with two detectors for the final measurement and the possibility of placing a third detector for the intermediate measurement. Finally, Miller repeated the argument for a system of tandem Mach–Zehnder interferometers. In all these cases the 'pre-selection only' situation is considered. It is unnatural to apply the time-symmetrised formalism to such cases. However, it must be possible. Thus, I need not show that the time-symmetrised formalism has an advantage over the standard formalism for describing these situations, but only that it is consistent with the predictions of the standard quantum theory.

In the standard approach to quantum theory the probability for the result of a measurement of C at time t is given by equation (1). The claim of all the proofs is that the counterfactual interpretation of the ABL rule yields a different result. In all cases the final measurement at time t_2 has two possible results which we signify as '1_f' and '2_f'. The suggested application of the ABL rule is as follows. The probability for the result c_j is:

$$P(C = c_j) \;=\; P(1_f)\cdot P(C = c_j|1_f) + P(2_f)\cdot P(C = c_j|2_f), \tag{15}$$

where $P(C = c_j|1_f)$ and $P(C = c_j|2_f)$ are the conditional probabilities given by the ABL formula, equation (2), and $P(1_f)$ and $P(2_f)$ are the probabilities for the results of the final measurement. In the proofs, the authors show that equation (15) is not valid and conclude that the ABL formula is not applicable to their example and therefore that it is not applicable in general.

I will argue that the error in calculating equality (15) is not in the conditional probabilities given by the ABL formula, but in the calculation of the probabilities $P(1_f)$ and $P(2_f)$ of the final measurement. In all three cases it was calculated on the assumption that *no* measurement took place at time t. This is the error: *one cannot make this assumption here* since then the discussion about the probability of the result of the measurement at time t is meaningless. Unperformed measurements have no results (Peres, 1978). Thus, there is no surprise that the value for the probability $P(C = c_j)$ obtained in this way comes out different from the value predicted by the quantum theory.

Straightforward calculations show that if one uses the formula (15) with the probabilities $P(1_f)$ and $P(2_f)$ calculated on the condition that the intermediate measurement has been performed, then the result is the same as predicted by the standard formalism of quantum theory. Consider, for example, the experiment suggested by Sharp and Shanks: consecutive spin measurements with the three directions in the same plane and with relative angles θ_{ab} and θ_{bc}. The probability for the final result 'up' is

$$P(1_f) = \cos^2(\theta_{ab}/2)\cos^2(\theta_{bc}/2) + \sin^2(\theta_{ab}/2)\sin^2(\theta_{bc}/2), \tag{16}$$

and the probability for the final result 'down' is

$$P(2_f) \ = \cos^2(\theta_{ab}/2)\sin^2(\theta_{bc}/2) + \sin^2(\theta_{ab}/2)\cos^2(\theta_{bc}/2). \tag{17}$$

The ABL formula yields

$$P(up|1_f) = \frac{\cos^2(\theta_{ab}/2)\cos^2(\theta_{bc}/2)}{\cos^2(\theta_{ab}/2)\cos^2(\theta_{bc}/2) + \sin^2(\theta_{ab}/2)\sin^2(\theta_{bc}/2)} \tag{18}$$

and

$$P(up|2_f) = \frac{\cos^2(\theta_{ab}/2)\sin^2(\theta_{bc}/2)}{\cos^2(\theta_{ab}/2)\sin^2(\theta_{bc}/2) + \sin^2(\theta_{ab}/2)\cos^2(\theta_{bc}/2)}. \tag{19}$$

Substituting all these equations into equation (15) we obtain

$$P(up) = \cos^2(\theta_{ab}/2). \tag{20}$$

This result coincides with the prediction of standard quantum theory. It is a straightforward exercise to show in the same way that no inconsistency arises in the examples of Cohen[7] and Miller either.

[7] In Cohen's example the measurement at time t_2 is not a complete measurement and therefore the ABL formula (2) is not applicable to this case. The analysis requires a generalisation of the ABL formula given in Vaidman (1998a).

I have shown that one can apply the time-symmetrised formalism, including the ABL formula, to analyse the examples which allegedly lead to contradictions in the inconsistency proofs. In my analysis there was nothing 'counterfactual'. The proofs, however, claimed to show that a 'counterfactual interpretation' of the ABL rule leads to contradiction. What I have shown is that the examples presented in the proofs do not correspond to counterfactual situations and this is why they cannot be analysed in a counterfactual way. The contradictions in the proofs arise from a logical error in taking together the statement 'no measurement has been performed at t' and a statement about the probability of a result of this measurement which requires that 'the measurement has been performed at t'.

Let me demonstrate how similar erroneous 'counterfactual' reasoning can lead to a contradiction in quantum theory even in cases when the ABL rule is not involved. Consider two consecutive measurements of σ_x performed on a spin-$\frac{1}{2}$ particle prepared in a state $|\uparrow_z\rangle$. Let us ask (using the language of Sharp and Shanks): what is the probability that these measurements *would have had* the results $\sigma_x(t_1) = \sigma_x(t_2) = 1$, given that no such measurements in fact took place? Each spin measurement, if performed separately, has probability $\frac{1}{2}$ for the result $\sigma_x = 1$. According to standard quantum theory the fact that in the actual world the measurement at t_1 has been performed and $\sigma_x(t_1) = 1$ has been obtained does not ensure that in a counterfactual world in which σ_x was not measured at t_1, but at a later time t_2, the result has to be $\sigma_x(t_2) = 1$. Rather, we still have probability $\frac{1}{2}$ for this result. Thus, Sharp and Shanks' counterfactual reasoning leads us to the erroneous result that the probability for the results $\sigma_x(t_1) = \sigma_x(t_2) = 1$ is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

6. Counterfactual Interpretations of the ABL Probability Rule[8]

In this section I shall consider three ways to interpret the 'counterfactual interpretation'. The first interpretation I cannot comprehend, but I have to discuss it since it has been proposed and used in the criticism of the time-symmetrised quantum theory. I believe that I understand the meaning of the second interpretation, but I shall argue that it is not appropriate for the problem which is discussed here. The last interpretation is the one I want to adopt and I shall present several arguments in its favour.

> *Interpretation* (*a*). Counterfactual probability as the probability of the result of a measurement which has not been performed.

Let me quote Sharp and Shanks:

> [...] for, conditionalizing upon specified results of measurements of M_I and M_F, there is no reason to assign the same values to the following probabilities: the

[8] This section, taken from Vaidman (1996a), is, in fact, a preliminary version of Section 3, taken from Vaidman (1997). I bring it here because it discusses several points in more detail and mostly because Kastner refers explicitly to the text of this section.

> probability that an intervening measurement of M had the result m^j given that such a measurement in fact took place, and the probability that intervening measurement *would have had* the result m^j given that no such intervening measurement of M in fact took place. In other words there is no reason to identify $\text{Prob}(M = m^j|\mathbf{E}_M[\psi_I^i, \psi_F^k])$ and $\text{Prob}(M = m^j|\mathbf{E}[\psi_I^i, \psi_F^k])$ (1993, p. 491).

I cannot comprehend the meaning of the probability for the result $M = m^j$ given that the measurement M has not taken place. As far as I can see $\text{Prob}(M = m^j|\mathbf{E}[\psi_I^i, \psi_F^k])$ has no physical meaning. Sharp and Shanks continue:

> (For a classical illustration, consider a drug which, if injected to facilitate a medical test at t, has an effect, starting shortly after the test and persisting past t_F, on the value of the tested variable. Suppose that it is unknown whether a test was conducted at t, but that a value for the tested variable is obtained at t_F. Using the value at t_F, we would estimate differently the value prior to t depending on whether we assume that a test did or did not take place at t.)

This might explain what they have in mind, but the argument does not hold good since in many situations there is no quantum mechanical counterpart to the classical case of 'the value [of a tested variable] prior to t'. In standard quantum theory *unperformed experiments have no results* (again, see Peres, 1978).

Cohen and Hiley partially acknowledge the problem, admitting that at least in the framework of the orthodox interpretation this is a meaningless concept:

> In other words we cannot necessarily assume that the ABL rule will yield the correct probabilities for what the results of the intermediate measurements *would* have been, *if* they had been carried out, in cases where these measurements have not *actually* been carried out. In fact, this sort of counterfactual retrodiction has no meaning in the orthodox (i.e., Bohrian) interpretation of quantum mechanics, although it can legitimately be discussed within the standard interpretation and within some other interpretations of quantum mechanics (see, for example, Bohm and Hiley [1993]) (1996, p. 3).

I fail to understand Interpretation (a) in any framework. Maybe, if we restrict ourselves to the cases in which the system at the intermediate time is in an eigenstate of the variable which we intended to measure (but which we did not measure), we can associate the probability 1 with such unperformed measurements. This is close to the idea of Cohen (1995) to consider counterfactuals in the restricted cases corresponding to consistent histories introduced by Griffith (1984). But, as far as I can see, interesting situations do not correspond to consistent histories, and therefore no novel (relative to classical theory) features of quantum theory can be seen in this way. It is possible that what Cohen and Hiley (1996) have in mind is the interpretation (b) which I shall discuss next.

> *Interpretation* (*b*). Counterfactual probability as the probability of the result of a measurement were it to be performed, based on the information about the world in which the measurement has not been performed.[9]

[9] This interpretation is equivalent to Definition (iv) of Section 3. Note the discussion following Definition (iv) which is relevant for Interpretation (b) but is not repeated here.

At time t_1 we preselect the state $|\Psi_1\rangle$. We do not perform any measurement at time t. We perform a measurement at time t_2 and find the state $|\Psi_2\rangle$. We ask, what would be the probability for the results of a measurement performed at time t in a world which is identical to the actual world at time t_1.

This is a meaningful concept, but I believe that it is not adequate for discussing pre- and post-selected quantum systems because it is explicitly asymmetric in time. The counterfactual world is identical to the actual world at time t_1 and might not be identical to it at time t_2.

> *Interpretation* (*c*). Counterfactual probability as the probability for the results of a measurement if it had been performed in the world 'closest' to the actual world.[10]

This is identical in form and spirit to the theory of counterfactuals of Bennett (1984), although the context of the pre- and post-selected quantum measurements is somewhat beyond what he considered. This interpretation is explicitly time-symmetric. The formulation of Interpretation (c) given above, however, does not specify it completely and I shall now explain what I mean (in particular, what I mean by the word 'closest').

I have to specify the concept of 'world'. There are many parts of the world which do not interact with the quantum system in question, so their states are irrelevant to the result of the measurement. In our discussion we might include all these irrelevant parts, or might not, without changing any of the conclusions. There are other aspects of the world which are certainly relevant to the measurement at time t, but we postulate that they should be disregarded. Everything which is connected to our decision to perform the measurement at time t and all the records of the result of that measurement are not considered. Clearly, the counterfactual world in which a certain measurement has been performed is different from an actual world in which, let us assume, no measurement has been performed at time t. The profound differences are both in the future where certain records exist or do not exist, and in the past which must be different since one history leads to performing the measurement at time t and another history leads to no measurement.[11] However, our decision to make the measurement is not connected to the quantum theory, which makes predictions about the result of that measurement. We want to limit ourselves to the discussion of the time symmetry of the quantum system. We do not consider here the question of the time symmetry of the entire world. Therefore, we exclude the external parts from our consideration.

What constitutes a description of a quantum system itself is also a very controversial subject. The reality of the Schrödinger wave, the existence or non-existence of hidden variables etc. are subjects of heated debates. However, everybody agrees that the collection of all results of measurements is a consistent (although maybe not complete) description of the quantum system. Thus,

[10] This interpretation corresponds to Definition (ii) and Definition (iii) of Section 3.

[11] If a random process chooses between the two possibilities, then the past before this process might be the same.

I propose the following definition:

> *Definition.* A world 'closest' to the actual world is a world in which all measurements (except the measurement at the time t if performed) have the same results as in the actual world.

This definition overcomes the common objection according to which one should not consider together statements about pre- and post-selected systems regarding different measurements at time t because these systems belong to different ensembles. The difference is in their quantum state at the time period between t and t_2.[12] Formally, the problem is solved by considering only results of measurements and not the quantum state. The justification of this step follows from the rules of the game: it is postulated that the quantum system is not disturbed during the periods of time (t_1, t) and (t, t_2). Therefore, it is postulated that no measurement on the system is performed during these periods of time. Since unperformed measurements have no results, the difference between the ensembles has no physical meaning in the discussed problem.

From the alternatives I have presented here, only Interpretation (c) is time-symmetric. This is the reason why I believe that it is the only reasonable candidate for analysing the (time-symmetric) problem of measurements performed between two other measurements.

7. Kastner's Readings of the ABL Rule

Kastner (1999) puts in quotation marks two possible readings of the ABL rule, 'non-counterfactual' and '*bona-fide* counterfactual' (1999, p. 239). As far as I can see the two readings are not different: the first is a clarification of the second. In the papers about TSQT the statements frequently appear in the compact form of the second reading and the first reading is the correct explanation of its meaning.

Kastner, however, distinguishes between the two readings. She claims that there is a 'quantifier ambiguity' in the ABL formula (2) (or her equation (1)). She proposes to add a parameter, C, indicating the variable which was actually measured at time t. Her equation (1″) corresponds to the following modification of (2) for pre-selection of $A = a$ and post-selection of $B = b$:[13]

$$P(o_j|a, b;C) = \frac{|\langle b|\mathbf{P}_{O=o_j}|a\rangle|^2}{\sum_i |\langle b|\mathbf{P}_{O=o_i}|a\rangle|^2}. \tag{21}$$

[12] If one adopts our backward evolving quantum state, one can add that the systems are also different due to the backward evolving state between t and t_1.

[13] There are two differences between Kastner's (1″) and my (21). First, the former is applicable only to measurements of non-degenerate observables, while the latter is more general and is applicable also to observables with degenerate eigenvalues such as the observables associated with 'the particle being in a particular box' of the 3-boxes example discussed by Kastner. Second, Kastner does not always follow the semantic convention used in this and other papers on TSQT according to which the eigenvalue of an operator denoted by a capital letter is signified by the same lower-case letter; here, my notation o_j corresponds to Kastner's x_j.

According to Kastner, for a counterfactual reading of the ABL rule this formula should yield the probability for $O = o_j$ given that C was *actually* measured for all O, including $O \neq C$. It is crucial to understand the exact meaning of Kastner's words 'C was *actually* measured' (see especially her footnote 3).

The first possible reading of Kastner is that C is related to the counterfactual world, for which the formula should yield probabilities for o_j. This reading is equivalent to Interpretation (a) of Section 6, which is meaningless. Indeed, in the framework of quantum theory observables usually do not possess values. There is no meaning for 'probability of a value', only for 'probability of a result of a measurement'. If it is postulated that O is not measured (since another variable, C, is measured instead), then it is meaningless to ask what is the probability for o_j. In other words, the question is what parameters are kept fixed when the counterfactual world is considered. Kastner's notation, $P(o_j|a, b;C)$, suggests that a, b, and C are kept fixed in the counterfactual world, but then there is no meaning for probability of o_j.

The second possible reading of Kastner is that C relates to the actual world and the formula is related to a counterfactual world, in which another variable, O, is measured. In this case it is not clear what is kept fixed in the counterfactual world. If a and b are kept fixed, then how can C be relevant? The question is about the counterfactual world which is specified *completely* by a and b, so the information about what has happened in another (the actual) world is irrelevant.

Finally, it might be that Kastner assumes some hidden variables which are kept fixed, i.e. identical for actual and counterfactual worlds. Then C is relevant because it characterises the hidden variables: they are such that, given the intermediate measurement C, the result b is obtained. This is a modification of Interpretation (iv) of Section 3 (Interpretation (b) of Section 6); in the latter no measurement is performed at time t in the actual world, i.e. the operator C is the identity I. While it is not immediately obvious that the ABL formula fails for Interpretation (iv) (it does fail as proved in Section 3), it is obvious that Kastner's modification of the ABL formula (21) is not appropriate: the right-hand side is a well-defined expression in the framework of quantum theory which does not depend on C, while it is assumed that the left-hand side depends on C.

As explained in Section 3, the failure of the ABL rule for hidden variables readings is not surprising since the whole concept of hidden variables, as we know it now, is time-asymmetric. Hidden variables supposedly allow us to *predict* the result of a measurement before the measurement is performed. In any way, there are no hidden variables in TSQT and, therefore, this failure does not represent a problem.

It might be that Kastner and others have been misled by the term 'element of reality'. The words suggest something 'ontological', but in TSQT 'element of reality' is a technical term which describes a situation in which a certain counterfactual statement is true: the result of a measurement (*if* performed!) is known with certainty (see Section 4 and Vaidman, 1996b). The only meaning of an element of reality 'the particle is in box A' quoted by Kastner is that 'if

searched in A it has to be there with probability 1', nothing more. After quoting this, Kastner writes (1999, p. 241): 'This usage clearly implies that the properties of being in box A or being in box B are considered as possessed by the same pre- and post-selected particle'. But I emphasise that the 'same' only means that the two-statevector at time t is fixed (i.e. the results of the pre- and the post-selection measurements are fixed). The counterfactual worlds corresponding to 'being in A' and 'being in B' for the 'same' particles are different: in one world the particle is searched for in A and in another it is searched for in B.

8. Kastner's Analysis of the Sharp and Shanks Proof

Kastner makes a distinction between two counterarguments to the Sharp and Shanks proof which I presented in the two preprints (Vaidman, 1996a, 1997). From my point of view there is just one argument presented in different forms in the two papers. The most relevant parts of the two papers appear here: Sections 2–5 (Vaidman, 1997) and Section 6 (Vaidman, 1996a). Let me state the relations between the statements presented in these sections. Interpretation (a) of Section 6, which I consider to be meaningless, is not introduced formally in the other preprint; I present there (fourth paragraph of Section 5) only a brief discussion why it should be rejected. The counterpart of Interpretation (b) of Section 6 is Definition (iv) of Section 3 and the counterpart of Interpretation (c) is Definition (ii). Definition (iii) of Section 3 is the formulation of Definition (ii) (or Interpretation (c)) when applied to the particular situation corresponding to the ABL scenario. The difference between the preprints, as Kastner correctly noticed, is that in Vaidman (1997) (which is the revision of Vaidman (1996a)) I do not focus on the possibility of a counterfactual with true antecedent. I still think that this possibility is a genuine property of quantum counterfactuals. However, I realised that many readers were confused by this point and, since it is not central, I decided that I can persuade people better without emphasising this property.

Kastner writes that my counterarguments lead to what she calls the 'non-counterfactual' interpretation of the ABL rule 'which is not under dispute'. She then proceeds with the analysis of the Sharp and Shanks argument focusing on 'a failure of cotenability between the background conditions, S, and the antecedent P'. It seems to me that this failure of cotenability is similar to my argument against the proof of Sharp and Shanks. They claimed that the counterfactual interpretation of the ABL rule leads to predictions different from that of quantum theory. I claimed that their counterfactual interpretation has a logical error and therefore their proof is incorrect.[14] Kastner shows that a part of the proof of Sharp and Shanks, expressed in the left-hand side of the logical relation (6) of her paper, is false. From a false logical statement one cannot claim to

[14] Note, in particular, the last paragraph of Section 5 in which I show how the Sharp and Shanks argument leads to a contradiction in a situation where the ABL rule is not involved.

calculate correctly the probability for a result of a measurement. So the problem is not with the ABL formula, as Sharp and Shanks claimed, but with the proof, as I claimed.

Thus, it seems that Kastner contradicts herself when she shows that the Sharp and Shanks calculation of the probability of a particular result in their example cannot be applied due to the failure of the cotenability condition Γ (her equation (4a)), but nevertheless continues with a 'detailed description of the steps employed in the Sharp and Shanks proof', the 'proof that the counterfactual interpretation of the ABL rule leads to predictions incompatible with quantum mechanics'. In what follows I shall analyse this 'detailed description' relating to the equations and the notation of Kastner's paper.

With her equation (6) she defines the framework of the Sharp and Shanks argument, and in her equations (7) and (8) she reproduces the Sharp and Shanks calculation of the probability for the result $C = c_1$. In particular, the calculation assumes S, i.e. that the systems in question belong to a pre- and post-selected ensemble M, the ensemble which can be obtained with high probability in an experiment, given the assumption $\neg P$, that no intermediate measurement of C has been performed.

Then, in a short paragraph, she presents *her* argument: she notes that in general it is not possible for both S (which was derived on the assumption $\neg P$) and P to be true. This is what she names as the failure of cotenability. And this is why equation (7) 'may well be false'. (Compare this with my argument according to which it is wrong to use S when the question of probability of $C = c_1$ is considered, because this question requires accepting P.)

Next, Kastner presents the Sharps and Shanks strategy in which they compare their calculation (7) with calculation (9)–(10), which does not assume S. She poses the question: is it really correct to make the comparison that Sharp and Shanks made? What follows is the definition of the problem (6′) in which S is not assumed, and the corresponding calculation (12)–(18). This calculation is a reproduction (in a more general form) of *my* calculation from Vaidman (1997) (see (15)–(20) of this paper).

At that point Kastner makes claims which I cannot understand. She says that this calculation 'establishes' that Sharp and Shanks made the correct comparison. She 'pinpoints my error': my claim that the assumption of Sharp and Shanks that no intervening measurement has occurred in the 'counterfactual world' is flawed. However, immediately following this assertion, she herself says that 'it might be objected that S' [which follows from this assumption] 'is not the appropriate statement of background conditions'. She proceeds with an indirect argument against this objection which includes the odd claim that an alternative, S', which she has *defined* as 'the correct background conditions obtained when P is true' (the paragraph following her equation (10)) is also not cotenable with P.

The last paragraph of Kastner's Section 2 leaves me with several options for understanding Kastner's interpretation of the Sharp and Shanks example. The sentence 'In view of the existence of actual results at t_2, such results are an

indelible part of the history of world *i* and cannot be disregarded' might suggest that the question is about a counterfactual world in which the measurement at time t is performed, but the results of the measurements at t_2 are nevertheless as in the actual world. (This seems to contradict Kastner's concept of non-cotenability of P and S, but in fact the probability for such situation in a real experiment is usually small but not zero.) For *this* question, equation (7), the counterfactual calculation of Sharp and Shanks using the ABL formula *is correct*; and the fact that it does not yield the value given in equation (10) is not a contradiction, because the latter corresponds to a different situation in which there is only pre-selection.

Another possible reading of Kastner's paragraph is that in the counterfactual world with the intermediate measurement the results of the measurement at time t_2 are different, but still the results at t_2 in the actual world are relevant for calculating probabilities for the results of measurement at time t, ($t < t_2$). One can imagine such a situation if there are hidden variables which control the results of measurements beyond the standard quantum formalism. I have discussed this possibility in connection with Definition (iv) of Section 3. Indeed, in this case, the ABL formula yields incorrect results, but this is not surprising since this situation is intrinsically time-asymmetric: the actual and the counterfactual worlds coincide in the past, but not in the future, relative to time t.

From private communications with Sharp and Shanks I understood that the main goal of their paper was to show exactly this, i.e. that 'the ABL rule did not have the implications for hidden variables interpretations of quantum mechanics that Albert *et al.* (1985) had claimed'.[15] Careful reading of Sharp and Shanks shows that they indeed focus on this limited issue. However, the title and their conclusions suggests criticism of TSQT in a much wider sense, and led their followers to attack all possible counterfactual interpretations in the framework of TSQT.

9. Kastner's Criticism of the Time-Symmetrised Counterfactuals

Kastner again distinguishes between two, equivalent from my point of view, definitions of time-symmetrised counterfactuals given in the two papers (Vaidman, 1996a, 1997). Kastner's Definition 1 is my Interpretation (c) of

[15] This is a correct criticism which, however, 'pushes at an open door'. The letter of Albert *et al.* (1985) indeed gives the impression that the authors undertake their discussion in the framework of the hidden variables theory. However, in their reply (Albert *et al.*, 1986) to the criticism of Bub and Brown (1986) they clearly stated that they do not (or, at least, they do not now) think that the results of their letter are applicable to hidden variable theories. Moreover, although this conclusion of Sharp and Shanks is correct, I still think that the alleged proof in their paper is flawed; I believe that I presented the correct proof of this point in Section 3, see the paragraph which includes equations (4) and (5).

Section 6 and her Definition 2 is my Definition (iii) of Section 3. I see a difference between the definitions only in phrasing and the generality of their applications.

An integral part of Definition 1 is the criterion that determines the 'closest' worlds: worlds with the same results of all measurements as in the actual world, except the measurements at time t. According to Definition 2 only the results of measurements at times t_1 and t_2 are the same. However, for the ABL-type situation discussed in Kastner's paper, in which at times t_1 and t_2 *complete* measurements are performed, the results of measurements performed outside the interval $[t_1, t_2]$ are irrelevant. Since it is also postulated that no other measurements are performed at the time intervals (t_1, t) and (t, t_2), the two definitions must lead to the same probabilities.

Let me start with an analysis of Kastner's criticism of Definition 1. She writes that in this definition I ignore the difference between the ensembles M and M' corresponding to the actual and the counterfactual worlds. The ensembles describe the results of measurements at t_1 and t_2. But according to my definition 'all measurements in a counterfactual world, excluding measurements at t, have the same results as in the actual world'. Therefore, it is postulated by the definition that there is *no* difference between the ensembles. There is nothing to ignore.

In her next paragraph, Kastner misunderstood the quote from my work about the difference which 'has no physical meaning'. The difference I have discussed is irrelevant here because it is related to the difference in the measurements (and results) performed at time t. The results of these measurements depend on, but do not (by fiat) influence, the results at t_1 and t_2.

Kastner proceeds with some calculations (equations (19)–(23)) which are related to 'an experiment along the lines of the Sharp and Shanks example'. She concludes her calculations by the claim that 'the probability of the existence of a closest possible world j as required by Definition 1 is extremely small'. But the 'closest possible world' is some hypothetical world which has to fulfill some requirements: it either exists or does not exist. There is no meaning for 'the probability of the existence' here in the same way as there is no meaning for the probability of the existence of a solution of a given mathematical problem. In a particular example that Kastner considers, the possible world which fulfills the requirements does exist and therefore, Definition 1 is legitimate.[16]

If we perform the experiment according to Kastner's set-up with large number of particles N twice, once with and another time without the intermediate measurements, then, indeed, there is a very small probability that the results at time t_2 will be identical in the two experiments. However, even if we perform

[16] In Kastner (1998) she claimed, based on the same derivation (equations (19)–(23)), that the probability of the existence of a possible world is not just small, but zero. Such a claim is meaningful and, if correct, makes my Definition 1 empty. However, her proof, as I showed in Vaidman (1998b, p. 18) was in error.

such an experiment twice without the intermediate measurements, the probability to obtain identical results is extremely small. Accidently, for Kastner's particular choice, $\theta_{ab} = \pi/2$, the probability is the same (it equals 2^{-N}) in both cases. What Kastner considers instead is the probability for obtaining identical weights of results 'up' in the post-selection measurement with and without the intermediate measurement. It is true that in two experiments without an intermediate measurement, the probability for obtaining identical weights is higher, but Kastner herself mentions in footnote 7 that the results themselves, and not just their statistical weights, are relevant. Note however, that as I stressed before, the analysis of any one of these probabilities is irrelevant because we apply our formalism only in the situations when the results of the measurement at t_2 *are* fixed.

Other misunderstandings appear in Kastner's criticism of Definition 2. First, she augments the two-statevector notation with a specification which observable has been measured at time t. This is against the whole idea of TSQT. The two-statevector is a complete description at time t in the sense that it yields probabilities for all possible measurements (and the weak values for weak measurements) at time t. The two-statevector is specified by the results of measurement at times *different* than t: the measurement at t does not have a direct influence on the two-statevector at t. See Vaidman (1998c) for careful review of this concept.

Kastner, instead, proposes her own definition of 'time-symmetrically fixed'. Although she writes that it is 'in the sense of Definition 2', her definition has nothing in common with my proposal. In my proposal there is no question: 'will the system have the same two-statevector?'. The two-statevector is given by fiat and this *is* the 'time-symmetric fixing'. The example which Kastner considers demonstrates how she distorts Definition 2. The meaning of 'the results of measurements performed on the system at times t_1 and t_2 are fixed' is that the results of the measurements at t_1 and at t_2 in actual and counterfactual worlds are the same. In Kastner's Fig. 4 this is not true. Only the results of the measurement at t_1 are the same.

It seems that behind Kastner's definition there is an idea of some kind of hidden variable: she discusses systems which 'would still have the same two-statevector' even if some different measurement were performed at t. To have the same two-statevector is, in the current context, to have the same result of a measurement at t_2 in a situation in which *a priori* it is not certain on the basis of quantum mechanical laws. Standard quantum theory does not ensure the *same* result at t_2 even if the same measurement is performed at t and the situation is different only because of a change in some unrelated variable. The same result at t_2 is not ensured also in the situation which Kastner analyses in Section 4. Her quote: 'Definition 2 will also be tenable for this case, since all appropriately pre-selected systems which are post-selected via no intervening measurement would also, with probability 1, be post-selected via an intervening measurement of either the pre- or post-selection observable' shows us again that she relies on hidden variables.

10. Conclusions

The definition of counterfactuals in quantum theory which I propose, Definition (ii) of Section 3, is very simple-minded. It seems to me that if one reads my definition as it is, without trying to find something beyond it,[17] then it is complete and unambiguous. I believe that the definition is helpful in resolving some controversies about quantum counterfactuals (see my attempts in this direction in Vaidman (1998d, 1999)).

In principle, a counterfactual statement such as Definition (iii) of Section 3 (Kastner's Definition 2) is testable in a laboratory by creating a large ensemble of systems with measurements at t_1, t, and t_2; choosing the subensemble (the pre- and post-selected ensemble) with fixed results at t_1 and t_2; choosing (out of this subensemble) the subensemble with a particular measurement at t; and finally by making a statistical analysis of the results of the measurement at t on this final subensemble. Because it is testable, philosophers might be reluctant to consider the construct which I define as counterfactual, in spite of the fact that formally it corresponds to the general form of counterfactual statements, Definition (i) of Section 3. This is a semantic issue. I distinguish in Section 4 between situations in which only a single statement of the form of Definition (iii) is considered and situations in which several such statements for different variables, all related to a single system, are considered. Since quantum theory does not allow simultaneous measurements of certain variables, in the latter situation the set of statements is, in fact, not testable.

Probably the simplest example of this kind is the set-up discussed by Kastner in which a free spin-$\frac{1}{2}$ particle is pre-selected at t_1 in a state with spin 'up' in one direction and post-selected at t_2 with the spin 'up' in another direction. At the intermediate time t there are two counterfactual statements (elements of reality according to my Definition (v)): indeed, the results of two, in general incompatible, measurements of spin components (in the pre- and post-selected directions) are certain. Consider now a rare quantum event in which a large number N of such identically pre- and post-selected particles undergo a *weak measurement* of a total spin in some other co-planar direction at an intermediate time t. A theorem from Aharonov and Vaidman (1991, p. 2325): *a weak measurement in a situation in which the result of a usual measurement is known with certainty yields the same result*, together with a general property of weak values, $(X + Y)_w = X_w + Y_w$, lead to a useful application of the counterfactuals. Counterfactual statements about unperformed experiments help us to find out the expected result of an actually performed experiment: the weak value of the total spin. In this time-symmetric example the statements about certain results of the spin measurements in the two directions are clearly counterfactual. I know that these measurements have not been performed because in the actual world the only interaction at time t is a weak coupling to a measuring device

[17] Apart from 'redefinition' of Definition 2, Kastner uses the word 'ontological' in her paper. TSQT does not make ontological claims. It is more a novel *formalism* than an *interpretation.*

which measures another observable. Moreover, the two measurements are incompatible, therefore the measurements could not have been performed together. I analyse more examples of such situations in Sections 8 and 9 of Vaidman (1997).

Kastner concludes that a 'counterfactual interpretation of the ABL rule is not valid in general'. I have shown that this conclusion follows from Kastner's particular reading of the word 'counterfactual'. I find that by and large Kastner supports my claim that counterfactuals, in a sense which differs from mine, are inconsistent in the framework of TSQT. She rejects my approach saying that this is a 'non-counterfactual' reading. I disagree about this semantic issue. More importantly, I disagree with Kastner's claim that in several articles in the framework of TSQT the 'bona fide' counterfactual reading of the ABL rule has been used (from which it would follow that, since this reading is inconsistent, the results of these articles, various 'curious' quantum effects, are wrong). This claim, however, was not proved, but only stated in Kastner's paper. Indeed, essentially only time-asymmetric examples were analysed in her paper.

I have shown that Kastner's criticism of my definitions of time-symmetrised counterfactuals is unfounded. In her discussion of Definition 1 she is misled by small probability of post-selection results corresponding to 'curious' quantum effects, while what is required for legitimacy of Definition 1 is only the possibility of such a result. Definition 2 (Definition (iii) of Section 3) she interprets in a particular sense. I agree that in this sense it 'has no clear physical meaning'. I suspect that she rejects the literal interpretation of Definition 2, the one which I adopt, because she views it as non-counterfactual.

Kastner finds 'an interesting special case in which the ABL rule may be correctly used in a counterfactual sense', the one which corresponds to *consistent histories* (Griffiths, 1984). First, if one adopts Kastner's interpretation of 'time-symmetrically fixed', this claim is problematic. As I explained at the end of the previous section, without introducing hidden variables it is simply false, and she has not defined hidden variables which will make it true. Second, I do not find the consistent histories approach fruitful in this context. In particular, it prescribes not to consider together incompatible families of histories (Griffiths, 1998). The two counterfactuals from the above example of the pre- and post-selected spin-$\frac{1}{2}$ particle about definite results of spin measurements in two directions (which can be derived also in the consistent histories framework) belong to two incompatible families and therefore they should not be considered together. Thus, in the consistent histories framework we cannot derive the result obtained above in the framework of TSQT regarding the intermediate weak measurement.

Many quantum mechanical effects are dramatically different from phenomena which can be explained classically. Language and philosophy which were developed during the time that no one suspected quantum phenomena have significant difficulties in defining and explaining quantum reality. This seems to be the reason for numerous controversies in this field. I believe

that discussing and resolving these controversies is of crucial importance for understanding our world.

Acknowledgements—It is a pleasure to thank Yakir Aharonov, David Albert, Avshalom Elitzur, Lior Goldenberg, Yoav Ben-Dov, Igal Kvart, Abner Shimony and Stephen Wiesner for helpful discussions and Willy De Baere, Jonathan Bennett, Jeremy Butterfield, Ruth Kastner, Philip Pearle, Niall Shanks, and David Sharp for useful correspondence. The research was supported in part by grant 471/98 of the Basic Research Foundation (administered by the Israel Academy of Sciences and Humanities).

References

Aharonov, Y. and Albert, D. Z. (1987) 'The Issue of Retrodiction in Bohm's Theory', in B. J. Hiley and F. D. Peat (eds) *Quantum Implications* (New York: Routledge and Kegan Paul), pp. 224–226.

Aharonov, Y., Albert, D. Z. and Vaidman, L. (1986) 'Measurement Process in Relativistic Quantum Theory', *Physical Review* **D34,** 1805–1813.

Aharonov, Y., Bergmann, P. G. and Lebowitz, J. L. (1964) 'Time Symmetry in the Quantum Process of Measurement', *Physical Review* **B134,** 1410–1416.

Aharonov, Y. and Vaidman, L. (1990) 'Properties of a Quantum System During the Time Interval Between Two Measurements', *Physical Review* **A41,** 11–20.

Aharonov, Y. and Vaidman, L. (1991) 'Complete Description of a Quantum System at a Given Time', *Journal of Physics* **A24,** 2315–2328.

Albert, D. Z., Aharonov, Y. and D'Amato, S. (1985) 'A Curious New Statistical Prediction of Quantum Theory', *Physical Review Letters* **54,** 5–7.

Albert, D. Z., Aharonov, Y. and D'Amato, S. (1986) 'Comment on 'Curious Properties of Quantum Ensembles which have been Both Pre-selected and Post-Selected', *Physical Review Letters* **56,** 2427.

Bedford, D. and Stapp, H. P. (1995) 'Bell's Theorem in an Indeterministic Universe', *Synthese* **102,** 139–164.

Bennett, J. (1984) 'Counterfactuals and Temporal Direction', *Philosophical Review* **93,** 57–91.

Bub, J. and Brown, H. (1986) 'Curious Properties of Quantum Ensembles which have been Both Pre-selected and Post-Selected', *Physical Review Letters* **56,** 2337–2340.

Cohen, O. (1995) 'Pre- and Post-Selected Quantum Systems, Counterfactual Measurements, and Consistent Histories', *Physical Review* **A51,** 4373–4380.

Cohen, O. and Hiley, B. J. (1996) 'Elements of Reality, Lorentz Invariance and the Product Rule', *Foundations of Physics* **26,** 1–15.

Ghirardi, G. C. and Grassi R. (1994) 'Outcome Predictions and Property Attribution: the EPR Argument Reconsidered', *Studies in History and Philosophy of Science* **25,** 397–423.

Griffiths, R. B. (1984) 'Consistent Histories and the Interpretation of Quantum Mechanics,' *Journal of Statistical Physics* **36,** 219–272.

Griffiths, R. B. (1998) 'Choice of Consistent Family, and Quantum Incompatibility,' *Physical Review* **A57,** 1604–1618.

Kastner, R. E. (1998) 'Time-symmetrized Quantum Theory, Counterfactuals, and 'Advanced Action', Los Alamos e-print archives quant-ph/9806002-v6.

Kastner, R. E. (1999) 'Time-Symmetrised Quantum Theory, Counterfactuals, and 'Advanced Action', *Studies in History and Philosophy of Science* **30,** 237–259.

Lewis, D. (1973) *Counterfactuals* (Oxford: Blackwells).

Lewis, D. (1986) 'Counterfactual Dependence and Time's Arrow' (reprinted from *Nous* **13,** 455–476 (1979)) and 'Postscripts to "Counterfactual Dependence and Time's Arrow" ', in D. Lewis, *Philosophical Papers*, *Vol. II* (Oxford: Oxford University Press), pp. 32–64.

Mermin, N. D. (1989) 'Can You Help Your Team Tonight by Watching on TV? More Experimental Metaphysics from Einstein, Podolsky, and Rosen', in J. T. Cushing and E. McMullin (eds), *Philosophical Consequences of Quantum Theory*: *Reflections on Bell's Theorem* (Notre Dame: University of Notre Dame Press), pp. 38–59.

Miller, D. J. (1996) 'Realism and Time Symmetry in Quantum Mechanics', *Physical Letters* **A222,** 31–36.

Penrose, R. (1994) *Shadows of the Mind* (Oxford: Oxford University Press).

Peres, A. (1978) 'Unperformed Experiments Have no Results', *American Journal of Physics* **46,** 745–747.

Peres, A. (1993) *Quantum Theory*: *Concepts and Methods* (Dordrecht: Kluwer Academic).

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

Redhead, M. L. G. (1987) *Incompleteness*, *Nonlocality*, *and Realism*: *a Prolegomenon to the Philosophy of Quantum Mechanics* (New York: Oxford University Press).

Sharp, W. D. and Shanks, N. (1993) 'The Rise and Fall of Time-Symmetrized Quantum Mechanics', *Philosophy of Science* **60,** 488–499.

Skyrms, B. (1982) 'Counterfactual Definiteness and Local Causation', *Philosophy of Science* **49,** 43–50.

Vaidman, L. (1993) 'Lorentz-Invariant "Elements of Reality" and the Joint Measurability of Commuting Observables', *Physical Review Letters* **70,** 3369–3372.

Vaidman, L. (1996a) 'Defending Time-Symmetrized Quantum Theory', Los Alamos e-print archives quant-ph/9609007.

Vaidman, L. (1996b) 'Weak-Measurement Elements of Reality', *Foundations of Physics* **26,** 895–906.

Vaidman, L. (1997) 'Time-Symmetrized Counterfactuals in Quantum Theory', Tel-Aviv University preprint TAUP 2459-97, Los Alamos e-print archives quant-ph/9807075.

Vaidman, L. (1998a) 'Validity of the Aharonov-Bergmann-Lebowitz rule', *Physical Review* **A57,** 2251–2253.

Vaidman, L. (1998b) 'Defending Time-Symmetrized Quantum Counterfactuals', Los Alamos e-print archives quant-ph/9811092.

Vaidman, L. (1998c) 'Time-Symmetrized Quantum Theory', *Fortschritte Der Physik* **46,** 729–739.

Vaidman, L. (1998d) 'Time-Symmetrized Counterfactuals in Quantum Theory', Los Alamos e-print archives quant-ph/9807042.

Vaidman, L. (1999) 'Variations on the Theme of the Greenberger–Horne–Zeilinger Proof', *Foundations of Physics* **29,** 615–630.

Vaidman, L., Aharonov, Y. and Albert, D. (1987) 'How to Ascertain the Values of σ_x, σ_y, and σ_z of a Spin-1/2 Particle', *Physical Review Letters* **58,** 1385–1387.