The Problem of Conscious Observation in Quantum Mechanical Description

H. D. Zeh

Institut für Theoretische Physik, Universität Heidelberg,
Philosophenweg 19, D-69120 Heidelberg, Germany.

Abstract

This is the retyped and slightly reformulated version of a paper that was circulated and semi-officially published already in 1981 in the Epistemological Letters of the Ferdinand-Gonseth Assosiation in Biel (Switzerland) as Letter No 63.0. (Therefore, terms such as "new" or "recent" refer to that year.) This paper offered a "discussion of the epistemological implications of quantum theory". In its Conclusion, it introduced the term "multi-consciousness interpretation" for a variant of the Everett interpretation. This has since been rediscovered several times (more or less independently), and become known as the "many-minds interpretation". Some comments and references have therefore been added at the end.

1 Introduction

John von Neumann seems to have first clearly pointed out the conceptual difficulties that arise when one attempts to formulate the physical process underlying subjective observation within quantum theory [1]. He emphasized the latter's incompatibility with a psycho-physical parallelism, the traditional way of reducing the act of observation to a physical process. Conventional descriptions, based on the assumption of a physical reality in space and time, either assume a "coupling" (causal relationship — one-way or bidirectional) of matter and mind, or disregard the whole problem by retreating to pure behaviorism. However, even this may remain problematic when one attempts to describe classical behavior in quantum mechanical terms. Neither position can be upheld without fundamental modifications in a consistent quantum mechanical description of the physical world.

The problems of formulating a process of observation within quantum theory arise because of quantum nonlocality (quantum correlations or "entanglement" as part of the generic state), which in turn may be derived as a consequence of the superposition principle. For dynamical reasons, this non-locality does not even approximately allow the physical state of a local system (such as the brain or parts thereof) to exist [2]. Hence, no state of the mind can exist "parallel" to it (that is, correspond to it one-to-one or determine it).

The problem does not only concern the philosophical issue of matter and mind. It has immediate bearing on quantum physics itself, as the state vector seems to suffer the well known reaction upon observation: its "collapse". For this reason Schrödinger even argued that the wave function might not represent a physical state (neither of the system itself, nor of a system carrying information about it), but should rather have a "fundamental psychic meaning".

This sitation appears so embarrassing to most physicists that many of them tried hard (not least in these Epistemological Letters) to find a local reality behind the formalism of quantum theory. For some time their effort was borne by the hope that quantum correlations could be understood as statistical correlations arising from an unknown ensemble interpretation of quantum theory. (An ensemble explanation *within* quantum theory can be excluded [2].) However, Bell's work has demonstrated quite rigorously that any local reality — regardless of whether it can be experimentally confirmed in principle or not — would necessarily be in conflict with some predictions of quantum theory. Less rigorous though still quite convincing

arguments were known before in the form of the dynamical completeness of the Schrödinger equation for describing isolated microscopic systems, in particular those containing quantum correlations (such as many-electron atoms).

Although the evidence in favor of quantum theory (and against local realism) now appears overwhelming, the continued search for a traditional solution may be understandable in view of the otherwise arising epistemological problems. On the other hand, in the absence of any empirical hint how to revise quantum theory, it may be wise to take the quantum description of physical reality in terms of non-local state vectors for granted, and consider its severe consequences seriously. Such an approach may be useful regardless of whether it will later turn out to be of limited validity.

The conventional ("Copenhagen") pragamatic attitude of switching between classical and quantum concepts by means of intuitive considerations does, of course, not represent a consistent description. It should be distinguished from that wave-particle duality which can be incorporated into the general concept of a state vector (namely, the occupation number representation for wave modes). Unfortunately, personal tendencies for local classical or for non-local quantum concepts to describe "true reality" seem to form the major source of misunderstandings between physicists — cf. the recent discussion between d'Espagnat and Weißkopf [4].

It appears evident that conscious awareness must in some way be coupled to local physical systems: our physical environment has to interact with and thereby influence our brains in order to be perceived. There is even convincing evidence supporting the idea that *all* states of awareness reflect physico-chemical processes in the brain. These neural processes are usually described by means of classical (that is, local) concepts. One may speculate about the details of this coupling on purely theoretical grounds [5], or search for them experimentally by performing neurological and phychological work. In fact, after a few decades of exorcizing consciousness from psychobiology by retreating to pure behaviorism, the demon now seems to have been allowed to return [6]. On closer inspection, however, the concept of consciousness *as used* turns out to be a purely behavioristic one: certain aspects of behavior (such as language) are rather conventionally associated with consciousness. It is indeed strictly impossible for epistemological reasons to derive the concept of *subjective* consiousness (awareness) as *emerging* from a physical ("outside") world. Nonetheless, subjectivity need not form an "epistemological impasse" (Pribram's term [7]), but to grasp it may require combined efforts from physics, psychology and epistemology.

2 The Epistemology of Consciousness

By inventing his malicious demon, Descartes demonstrated the impossibility of *proving* the reality of the observed (physical) world. This hypothetical demon, assumed to delude our senses, may thereby be thought of as part of (another) reality — similar to an indirect proof.

On the other hand, Descartes' even more famous *cogito ergo sum* is based on our conviction that the existence of subjective sensations cannot be reasonably doubted. Instead of forming an epistemological impasse, subjectivity should thus be regarded as an epistemological gateway to reality.

Descartes' demon does not disprove a real physical world — nor does any other epistemological argument. Rather does it open up the possibility for a *hypothetical* realism, for example in the sense of Vaihinger's *heuristic fictions* [8]. Aside from having to be intrinsically consistent, this hypothetical reality has to agree with observations (perceptions), and describe them in the most economic manner. If, in a quantum world, the relation between (ultimately subjective) observations and postulated reality should turn out to differ from its classical form (as has often been suggested for reasons of consistency), new non-trivial insights may be obtained.

While according to Descartes my own sensations are beyond doubt to me, I cannot *prove* other people's consciousness even when I presume their physical reality. (This was the reason for eliminating it from behavioristic psychology.) However, I may better (that is, more economically) "understand" or predict others' behavior (which I *seem* to observe in reality) if I *assume* that they experience similar sensations as I do. In this sense, consciousness (beyond solipsism) is a heuristic concept precisely as reality. *There is no better epistemological reason to exorcise from science the concept of consciousness than that of an external reality.*

A consequence of this heuristic epistemological construction of physical *and* psychic reality is, of course, that language may give information about the content of others' consciousness. This argument emphasizes the epistemologically derived (rather than dynamically emerged) nature of this concept. However, only that part of others' consciousness can be investigated that manifests itself as some form of behavior (such as language). For this reason it may indeed be appropriate to avoid any fundamental concept of consciousness in psychobiology. This requires that conscious behavior (behavior as though being conscious) can be completely explained as emerging — certainly a meaningful conjecture. It would have to include our private (subjectively experienced) consciousness if a psycho-physical paral-

lelism could be established. Only for such a dynamically passive parallelism would the physical world form a closed system that in principle allowed complete reductionism.

Before the advent of quantum theory this ivory tower position of physics could be upheld without posing problems. If, on the other hand, the nonlocal quantum concepts describe *real* aspects of the physical world (that is, if they are truly heuristic concepts), the parallelism has to be modified in some way. Such a modification may some day even turn out to be important in experimental psychology. It will be irrelevant wherever nonlocality can be neglected, as in present-day computers or most neural processes. However, the quasi-classical activities of neurons could be almost as far from consciousness as an image on the retina. The concept of "wholeness" — often emphasized as being important for complex systems such as the brain — is usually insufficiently understood: in quantum theory it is neither a mere dynamical wholeness (that is, an efficient interaction between all parts) nor is it restricted to the system itself. Dynamical arguments require it to be a *kinematical wholeness of the entire universe* (when regarded as composed of *spatial* parts) [2]. It may be neglected for certain ("classical") aspects only — not for a complete miscroscopic description that may be relevant for subjective perceptions.

3 Observing in a Quantum World

One possible consequence of these problems that inevitably arise in quantum theory would be to abandon the heuristic and generally applicable concept of a physical reality — explicitly [9] or tacitly. This suggestion includes the usual restriction to formal rules when calculating probability distributions of *presumed* classical variables in situations which are intuitively understood as "measurements" (but insufficiently or even inconsistently distinguished from normal "dynamical" interactions). Clearly, no general description of physical processes underlying awareness can be given in the absence of a physical reality, even though macroscopic behavior (including the dynamics of neural systems) can be described by means of the usual pragmatic scheme. This is quite unsatisfactory, since subjective awareness has most elementary meaning without *external* observation (as it would be required in the Copenhagen interpretation). Epistemologically, any concept of observation must be ultimately based on an observing subject.

It has been readily explained [2, 10] that this "non-concept" of aban-

doning any microscopic reality is not at all required. It is instead possible to interpret the "actual" state vector as representing this reality, since it may *act* (often as a whole) on what is observed. Moreover, in view of Bell's analysis of the consequences of quantum nonlocality it appears difficult to see what could possibly be gained from inventing novel fundamental concepts (hidden variables) without any empirical support. Thereby, two different solutions of the measurement problem appear possible: von Neumann's collapse or Everett's multi-universes interpretation [11]. In both cases a (suitably modified) psycho-physical parallelism can be re-established.

A dynamical collapse of the wave function would require nonlinear and nonunitary terms in the Schrödinger equation [12]. They may be extremely small, and thus become effective only through practically irreversible amplification processes occurring during measurement-like events. The superposition principle would then be valid only in a linearized version of the theory — perhaps related to wave function renormalization. While this suggestion may in principle explain quantum measurements, it would not be able to describe definite states of concsiousness unless the parallelism were artificially restricted to quasi-classical variables in the brain. Since nonlinear terms in the Schrödinger equation must lead to observable deviations from conventional quantum theory, they should at present be disregarded for similar reasons as hidden variables. Any suggested violation of the superposition principle must be viewed with great suspicion because of the latter's great and general success. For example, even superpositions of different vacua have proven heuristic (that is, to possess predictive power) in quantum field theory.

The problems thus arising when physical states representing consciousness are described within wave mechanics by means of nonlinear dynamical terms could possibly be avoided if these nonlinearities were themselves *caused* by consciousness. This has in fact been suggested as a way to incorporate a genuine concept of free will into the theory [13], but would be in conflict with the hypothesis of a closed physical description of the world.

If the Schrödinger equation is instead assumed to be universal and exact, superpositions of states of the brain representing different contents of consciousness are as unavoidable as Schrödinger's superposition of a dead and alive cat. However, because of unavoidable interaction with the environment, each component must then be quantum correlated with a different (almost orthogonal) state of the rest of the universe. This consequence, together with the way how we perceive the world, leads obviously to a "many-

worlds" interpretation of the wave function.[1] Unfortunately this name is misleading. The quantum world (described by a wave function) would correspond to *one* superposition of myriads of components representing *classically* different worlds. They are all dynamically coupled (hence "actual"), and they may in principle (re)combine as well as branch. It is not the quantum world that branches in this picture, but consciousness (or rather the state of its physical carrier), and with it the observed (apparent) world [2]. Only empirical knowledge thus indicates that consciousness is physically determined by (factor) wave functions in certain *components* of the total wave function.[2] The *existence* of "other" components (with their separate conscious versions of ourselves) is a heuristic fiction, based on the assumption of a general validity of dynamical laws that have always been confirmed when tested. When applied to classical laws and concepts, the analogous assumptions lead to the conventional model of reality in space and time. In the quantum mechanical model, *a collapse would represent a new kind of solipsism*, since it denies the existence of the practically unobservable.

Everett related his branching to the practically irreversible dynamical decoupling of components that occurs when microscopic properties are amplified to the macroscopic scale. This irreversibility requires specific initial conditions for the global state vector [5]. Such initial conditions will then, for example, also cause a sugar molecule to permanently send *retarded* "information" about its handedness into the universe by scattering photons and molecules. In this way, their relative phases become nonlocal, and thus cannot effect the physical states of local conscious observers (such as those of brains) any more. The separation of these components is dynamically "robust". There is no precise localization of the branch cut (while a genuine dynamical collapse would have to be *specified* as a dynamical law).

Nonetheless, Everett's branching in terms of quasi-classical properties does *not* appear sufficient to formulate a psycho-physical parallelism. Neither would this branching produce a definite factor state for some relevant part of the brain, nor does every decoherence process somewhere in the uni-

[1] Everett [11] suggested "branching" wave functions in order to discuss cosmology in strictly quantum mechanical terms (without an external observer or a collapse). I was later led to similar conclusions as a consequence of unavoidable quantum entanglement [2] — initially knowing neither of Everett's nor of Bell's work.

[2] It would always be possible to introduce entirely arbitrary unobservable ("hidden") variables as a hypothetical link between the wave function and consciousness. Given their (hypothetical) dynamics, the required quantum probabilities can then be postulated by means of their initial distribution. An example are the classical variables in Bohm's pilot wave theory [14].

verse describe conscious observation. Even *within* a robust branch, most parts of the brain will remain strongly quantum correlated with one another and with their environment.

Everett's branchings represent objective measurements — not conscious observations. A parallelism would require a far more fine-grained branching (from a local point of view) than that describing measurements, since it should correspond one-one to subjective awareness. The conjecture here is: does the (not necessarily robust) branching that is required for defining a parallelism then readily *justify* Everett's (apparently objective) branching into quasi-classical worlds?

The branching of the global state vector Ψ with respect to *two* different conscious observers (A and B, say) may be written in their Schmidt-canonical forms [5],

$$\Psi = \sum_{n_A} c^A_{n_A} \chi^A_{n_A} \phi^A_{n_A} = \sum_{n_B} c^B_{n_B} \chi^B_{n_B} \phi^B_{n_B} \quad , \tag{1}$$

where $\chi^{A,B}$ are states of the respective physical carriers of consciousness (presumably small but not necessarily local parts of the central nervous system), while $\phi^{A,B}$ are different states of the rest of the universe correlated to them. In order to describe the macroscopic *behavior* of (human) observers, one has to consider the analogous representation with respect to the states $\tilde{\chi}$ of their whole bodies (or relevant parts thereof),

$$\Psi = \sum_{k_A} \tilde{c}^A_{k_A} \tilde{\chi}^A_{k_A} \tilde{\phi}^A_{k_A} = \sum_{k_B} \tilde{c}^B_{k_B} \tilde{\chi}^B_{k_B} \tilde{\phi}^B_{k_B} \quad . \tag{2}$$

In particular, the central nervous system may be assumed to possess (usually unconscious) "memory states" (labelled by m_A and m_B, say) which are similarly robust under decoherence as the handedness of a sugar molecule. Time-directed *quantum causality* (based on the initial condition for the global wave function) will then force the Schmidt states $\tilde{\chi}^A$ and $\tilde{\chi}^B$ to approximately factorize in terms of these memory states [15],

$$\Psi \Rightarrow \sum_{m_A \mu_A} \tilde{c}^A_{m_A \mu_A} \tilde{\chi}^A_{m_A \mu_A} \tilde{\phi}^A_{m_A \mu_A} \approx \sum_{m_B \mu_B} \tilde{c}^B_{m_B \mu_B} \tilde{\chi}^B_{m_B \mu_B} \tilde{\phi}^B_{m_B \mu_B} \quad , \tag{3}$$

where μ_A and μ_B are additional quantum numbers. The "rest of the universe" thus serves as a sink for phase relations.

In general, the robust quantum numbers m_A and m_B will be partly correlated — either because of special interactions between the oberservers

(communication), or since they have arisen from the same cause (that is, from observations of the same object). These correlations define the concept of objectivization in quantum mechanical terms.

The genuine carriers of consciousness (described by the states χ in (1)) must *not* in general be expected to represent memory states, as there do not seem to be permanent contents of consciousness. However, since they may be assumed to interact directly with the rest of the $\tilde{\chi}$-system only, and since phase relations between different quantum numbers m_A or m_B would immediately become nonlocal, memory appears "classical" to the conscious observer. Each robust branch in (2), hence also each m-value, describes essentially a *separate* partial sum of type (1) when observed [16]. The emprirically relevant probability interpretation in terms of quasi-classical branches (including pointer positions) may, therefore, be derived from a similar (but fundamental) one for the *subjective branching* (with respect to all observers) that according to this interpretation defines the novel psycho-physical parallelism.

As mentioned before, macroscopic *behavior* (including behavior as though being conscious) could also be described by means of the pragmatic (probalistic) rules of quantum theory. An exact Schrödinger equation does not imply deterministic behavior of conscious beings, since one has to expect that macroscopic stimuli have to have microscopic effects in the brain before they cause macroscopic behavior. Thereby, interaction with the environment will intervene. Everett's "relative state" decomposition (1) with respect to the subjective observer state χ may then considerably differ from the objectivized branching (3), that would be meaningful with respect to all conceivable "external" observations. This situation may help to put definite meaning into Bohr's vague concept of *complementarity*.

4 Conclusion

The multi-universes interpretation of quantum theory (which should rather be called a *multi-consciousnesses interpretation*) seems to be the only interpretation of a universal quantum theory (with an exact Schrödinger equation) that is compatible with the way the world is perceived. However, because of quantum nonlocality it requires an appropriate modification of the traditional epistemological postulate of a psycho-physical parallelism.

In this interpretation, the physical world is described by Everett's wave function that evolves deterministically (Laplacean). This global quantum

state then defines an indeterministic (hence "branching") succession of states for all observers. Therefore, the world itself *appears* indeterministic — subjective in principle, but largely objectivized through quantum correlations (entanglement).

This quite general scheme to describe the empirical world is conceptually consistent (even though the parallelism remains vaguely defined), while it is based on the presently best founded physical concepts. The latter may some day turn out to be insufficient, but it is hard to see how any future theory that contains quantum theory in some approximation may avoid similar epistemological problems. These problems arise from the contrast between quantum nonlocality (demonstrated by Bell's analysis to be part of *reality*) and the locality of consciousness "somewhere in the brain". Quantum concepts should be better founded than classical ones for approaching these problems.

5 Addendum of 1999

The above-presented paper of 1981 has here been rewritten as an e-print (with minor changes in formulations), since the solution of the quantum mechanical measurement problem proposed therein has recently gained interest, while the Epistemological Letters are now hard to access. The dynamical dislocalization of phase relations used in this article (and based on [2, 15]) has since become better known as *decoherence* (see [17]), while the "multi-consciousness interpretation" mentioned in the Conclusion has been rediscovered on several occasions. It is usually discussed as a "many-minds interpretation" [18, 19, 20, 21], but has also been called a "many-views" [22] or "many-perceptions" interpretation [23].

The conjectured quasi-classical nature of those dynamical states of neurons in the brain which can be observed "from outside" has recently been quantitatively confirmed by means of decoherence in an important paper by Tegmark [24]. To most of these states, however, the true physical carrier of consciousness somewhere in the brain may still represent an *external* observer system, with whom they have to interact in order to be perceived. Regardless of whether the ultimate observer systems are quasi-classical or possess essential quantum aspects, consciousness can only be related to factor states appearing in *branches* (components) of the global wave function if the Schrödinger equation is exact. Environmental decoherence represents *entanglement*, while *ensembles* of various (unpredictable but only individu-

ally real) outcomes would require a dynamical collapse of the wave function (that has never been observed).

An essential role of the mind for the occurrence of fundamental (though objective) quantum events was obviously assumed already by Heisenberg in his early "idealistic" interpretation of *a particle trajectory coming into being by our act of observing it*. Bohr, in his Copenhagen interpretation, insisted instead that classical outcomes arise in the apparatus during irreversible *measurements*, which he assumed not to be dynamically analyzable in terms of a microscopic reality. This link in the chain of interactions which form the observation of a quantum system can now be identified with the (first) occurrence of decoherence in this chain (described as a unitary but practically irreversible *dynamical* process — cf. [25]).

However, Bohr's approach as well as Heisenberg's uncertainty relations were meant to establish *bounds to a rational description of Nature*. (The popular simplistic view of quantum theory as merely defining stochastic dynamics in an otherwise classical world leads to the well known wealth of "paradoxes", which have all been *derived from a superposition principle that applies to all of reality*, that is, from an entangled global wave function.) Von Neumann's interpretation, on the other hand, is somewhat obscured by his use of observables, which should have no *fundamental* place in a theory of interacting wave functions. His postulate of a dynamical collapse representing conscious observations was later elaborated upon by London and Bauer [26], while Wigner [13] suggested an active influence of the mind on the physical (quantum) state. The latter would not have to affect objectively measurable probabilities. Stapp [21] has expressed varying views on this problem, while Penrose [27] speculated that human thinking, in contrast to classical computers, requires genuine quantum aspects (including entangled states *and* the collapse of the wave function).[3]

The Everett interpretation leads to "extravagant" consequences, because it does not invent any unobserved laws or variables or irrational elements in order to avoid them. Lockwood [19] is quite correct when he points out the essential role of decoherence for the many-minds interpretation. Unavoidable "continuous measurement" of all macroscopic systems by their

[3] There seems to be a certain confusion in the literature between logical *statements* (tautologies), which have no intrinsic relation whatsoever to the concept of time, and algorithmic *procedures* (in time) used to prove them. (Undecidable formal statements are meaningless, and hence not applicable to reality.) Similarly, a dynamical collapse of the wave function must not be regarded as "logic". This situation is reminiscent of the philosophical confusion of the concepts of *cause* and *reason*.

environments (inducing strong entanglement) was indeed initially discussed [2] precisely in order to support the concept of a universal wave function, in which "branching components" can only be separately experienced.

References

[1] J. von Neumann, *Mathematische Grundlagen der Quantentheorie* (Berlin 1932).

[2] H.D. Zeh, Found. Phys. **1**, 69 (1970).

[3] J.S. Bell, Physics **1**, 195 (1964).

[4] Scient. Amer. **242**, no. 5 and 8 (1980).

[5] H.D. Zeh, Found. Phys. **9**, 803 (1979).

[6] See, e.g., J.M. Davidson and R.J. Davidson (edts.) *The Psychobiology of Consciousness* (Plenum, New York 1980).

[7] K. Pribram, in Ref. [6], p. 61.

[8] H. Vaihinger, *Die Philosophie des Als Ob* (Berlin 1911).

[9] W. Pauli's letters to M. Born in: M. Born, *Briefwechsel zwischen A. Einstein und M. Born* (Nymphenburger, München 1969).

[10] H.D. Zeh, Epist. Lett. no. 49.0 (1980).

[11] H. Everett, Rev. Mod. Phys. **29**, 454 (1957).

[12] Ph. Pearle, Phys. Rev. **D13**, 857 (1976).

[13] E. Wigner, in: L.J. Good (edt.), *The Scientist Speculates* (Heinemann, London 1962).

[14] D. Bohm, Phys. Rev. **85**, 166 (1952); J.S. Bell, Epist. Lett. no. 37.0 (1978) — also published in: C. Isham, R. Penrose, and D. Sciama (edts.), *Quantum Gravity 2* (Clarendon Press, Oxford 1981).

[15] H.D. Zeh, in: B. d'Espagnat (edt.), *Foundations of Quantum Mechanics*, Enrico Fermi School of Physics **IL**, (Academic, New York 1972).

[16] H.D. Zeh, Found. Phys. **3**, 109 (1973).

[17] D. Giulini, E. Joos, C. Kiefer, J. Kupsch, I.-O. Stamatescu, and H.D. Zeh, *Decoherence and the Appearance of a Classical World* (Springer, Berlin 1996).

[18] D. Albert and B. Loewer, Synthese **77**, 195 (1988).

[19] M. Lockwood, *Mind, Brain and the Quantum: The Compound 'I'* (Basil Blackwell, Oxford, 1989); Brit. J. Phil. Sci. **47**, 159 (1996).

[20] M.J. Donald, Found. Phys. **22**, 1111 (1992); **25**, 529 (1995).

[21] H. Stapp, in: P. Lahti and P. Mittelstaedt (edts.), *Symposium on the Foundation of Modern Physics* (World Scientific, Singapore 1991); *Mind, Matter, and Quantum Mechanics* (Springer, Berlin 1993).

[22] E. Squires, Eur. J. Phys. **8**, 171 (1987); *Conscious Mind in the Physical World* (Hilger, Bristol 1990).

[23] D.N. Page, e-print quant-ph/9506010.

[24] M. Tegmark, e-print quant-ph/9907009. To be published in Phys. Rev. **E**.

[25] H.D. Zeh, *The Physical Basis of the Direction of Time*, 3^{rd} edn. (Springer, Berlin 1999).

[26] F. London and E. Bauer, *La théorie d'observation en Méchanic Quantique* (Hermann, Paris 1939).

[27] R. Penrose, *Shadows of the Mind* (OUP, Oxford 1994).