Mind-body interaction and modern physics

Charis Anastopoulos*

Department of Physics, University of Patras, 26500 Greece

March 12, 2020

Abstract

The idea that mind and body are distinct entities that interact is often claimed to be incompatible with physics. The aim of this paper is to disprove this claim. To this end, we construct a broad mathematical framework that describes theories with mind-body interaction (MBI) as an extension of current physical theories. We employ histories theory, i.e., a formulation of physical theories in which a physical system is described in terms of (i) a set of propositions about possible evolutions of the system and (ii) a probability assignment to such propositions. The notion of dynamics is incorporated into the probability rule. As this formulation emphasises logical and probabilistic concepts, it is ontologically neutral. It can be used to describe mental 'degrees of freedom' in addition to physical ones. This results into a mathematical framework for psycho-physical interaction ($\Psi\Phi$I formalism). Interestingly, a class of $\Psi\Phi$I theories turns out to be compatible with energy conservation.

1 Introduction

Any theory of mind that postulates mental states as fundamentally different from physical states has to explain how mental states can influence physical states and vice versa. The existence of such a *mind-body interaction* (MBI) is a natural common-sense belief. However, no theory has been developed to substantiate this belief, and several arguments have been put forward that such interactions are incompatible with the physical sciences.

*anastop@upatras.gr

In this paper, we contend the opposite: the conceptual tools for describing MBI are already present in contemporary theoretical physics. MBI is not contained within *existing* fundamental theories of physics, but those theories have a *mathematical structure* that can be employed in order to formulate theories with MBI. We demonstrate this by the explicit construction of a general mathematical framework for such theories.

Contemporary physics is based on a small number of fundamental theories—like quantum theory and general relativity—that pertain to describe all phenomena in their domain. These theories are mathematical: their concepts are represented by mathematical objects, and their principles can be expressed as mathematical axioms. Physics theories place a strong emphasis on the causal and structural features of the objects they study, and give relatively little attention to their ontology.

The emphasis on mathematical structure is a crucial factor that could lead to theories of MBI that make empirically testable predictions, even if their ontology is vague. As a matter of fact, our current fundamental theory for the microcosm, quantum mechanics, works exactly this way. To quote Dirac [1],

When you ask what are electrons and protons I ought to answer that this question is not a profitable one to ask and does not really have a meaning. The important thing about electrons and protons is not what they are but how they behave, how they move. I can describe the situation by comparing it to the game of chess. In chess, we have various chessmen, kings, knights, pawns and so on. If you ask what a chessman is, the answer would be that it is a piece of wood, or a piece of ivory, or perhaps just a sign written on paper, or anything whatever. It does not matter. Each chessman has a characteristic way of moving and this is all that matters about it. The whole game of chess follows from this way of moving the various chessmen.

We find the attitude expressed by Dirac to be very fruitful for the mind-body problem. One should simply substitute the "electrons and protons" in Dirac's quote with thoughts, qualia or any other mental object. We can even expand on Dirac's metaphor. A theory with MBI is analogous to a game that is played with two chess-boards, the mental board and the physical board. We should look for a new set of rules that describe how the movements of the pieces on one board are interconnected and affect the movements of the pieces on the other.

To this end, we must first reformulate the existing physical theories in a way that is amenable to such a generalization. We will show that the logical reformulation of physical theories in terms of histories—see, Sec. 3 for explanations of the terminology—satisfy this condition. The result is general mathematical framework for theories that describe MBI. We will refer to this framework as *ΨΦI formalism*, i.e., a formalism for theories with psycho-physical interaction.

Such theories will involve psycho-physical laws that 'relate experience to elements of physical theory' as proposed by Chalmers [2,3]. However, in contrast to Chalmers' proposal, the ΨΦI formalism does not describe the physical world as *causally closed*. The physical and the mental fully act upon each other. It turns out that some ΨΦI theories are fully compatible with energy conservation, hence removing a common objection to MBI.

The structure of this paper is the following. In Sec. 2, we explain how interactions of fundamentally different substances are described in current theories of physics. In Sec. 3, we present the main ideas of histories-based reformulations of physical theories. In Sec. 4, we argue that the natural way to explain the physical correlates of consciousness is to view such correlations as dynamical, i.e., as arising from the interaction of mental and physical degrees of freedom. Then, we show how this interaction is described in the ΨΦI formalism. In Sec. 5, we discuss the role of energy and information in the ΨΦI formalism and how they relate to open problems in physics. In the last section, we summarise and discuss our results.

2 General Relativity as a prototype for MBI

A traditional objection to theories of MBI is that if mental and physical properties are ontologically different, then they lack the communality that is necessary for interaction. This objection was raised against Cartesian dualism already in the 17th century. It was a strong argument as long as physical interactions were deemed to be mechanistic and to be based on physical contact of extended objects. Today, this argument has little force. First, because Bell's theorem asserts that it is impossible to describe the fundamental physical interactions mechanistically. Second, because physics already contains theories about the interaction between radically different 'substances'. General Relativity (GR), our current theory of gravity, describes the interaction between spacetime geometry and matter. In this

section, we elaborate on the latter point, and we argue that GR provides a methodological template for formulating theories of MBI.

We first explain the meaning of the terms "spacetime geometry" and "matter". Spacetime geometry is the structure that determines what clocks and rods measure, i.e., spatial and temporal distances between physical events. Events are represented by points on a four-dimensional manifold M. Three coordinates on M refer to space and one coordinate refers to time. In GR, matter can be taken roughly to refer to entities that extend in space and carry energy and momentum. Spacetime geometry and matter are essentially distinct, and their differences are not blurred by the fact that they interact.

General Relativity describes the interaction between matter and geometry by embedding both entities within a broad mathematical framework, namely, *Lagrangian field theory* (LFT) [4]. The LFT was initially conceived as a generalization of classical mechanics for continuous systems. The properties of continuous systems are expressed in terms of *classical fields*. A classical field is a map $\phi : M \to S$, for some set S, i.e., a maps that assigns one mathematical object $\phi(X) \in S$ to each spacetime point X. LFT enables us to describe a system's dynamics in terms of partial differential equations that are satisfied by the classical fields.

Note that the formulation of GR in terms of LFT was not accidental. LFT originates from the tradition of 19th century's analytical dynamics. The latter was explicitly championed as an abstract theory of dynamics that could be used without commitment to a specific ontology of the microscopic structure of matter and/or the ether [5]. This is the reason why it survived the demise of mechanistic models for matter and ether, and it remains an indispensable physics tool.

Spacetime geometry is expressed in terms of a field g, the *Lorentzian metric*. The metric g incorporates all geometric information in a compressed form. Spatial and temporal distances are obtained by decompressing the information contained in the metric through the solution of the so-called *geodesic equations*. The correspondence between geometries and metrics is not one-to-one. A metric also carries some non-geometric information about the choice of a coordinate system, with the result that one geometry corresponds to an

infinity of different metrics[1].

In contrast, there is a huge loss of information when we use classical fields to describe matter. Fundamentally, matter is described by quantum theory; in the LFT description, quantum effects are ignored (or averaged out). The treatment of matter as continuous at macroscopic scales also implies that the discrete structure at the atomic level is ignored.

The description of matter and geometry in terms of classical fields is suboptimal for both matter and gravity. The field description of geometry contains too much information, the field description of matter contains too little. The LFT formulation of matter-gravity interaction in GR is a working compromise, and not a perfect fit. Nonetheless, it suffices for a formulation of a theory in which 'spacetime tells matter how to move; matter tells spacetime how to curve' [6].

The compromises involved in the formulation of GR are anything but benign. They are the source of major problems in any attempt to extend GR, for example, towards a quantum theory of gravity[2]. Nonetheless, GR works. It has a consistent and elegant mathematical structure, and it leads to predictions with excellent agreement to the experiment.

GR suggests the following strategy for describing the interaction between two ontologically different entities A and B.

1. Identify an appropriate mathematical framework Dyn for the description of dynamics (the analogue of LFT for GR).

2. Represent entity A by mathematical objects F_A and entity B by mathematical objects F_B, where both F_A and F_B fit within the structure of Dyn. These representations need not be one-to-one; they may involve either redundancy or information loss.

3. Identify all possible dynamics in Dyn that involve interaction between F_A and F_B.

In the present context, we want A to correspond to mental states/processes and B to physical states/processes. We have a very good idea of the mathematical structures involved in B, and, thus, of possible frameworks Dyn compatible with B. In the next

[1]The space of all Lorentzian metrics on the spacetime manifold M is denoted by $\mathrm{LRiem}(M)$. The space of spacetime geometries is the quotient manifold $\mathrm{LRiem}(M)/\mathrm{Diff}(M)$ where $\mathrm{Diff}(M)$ is the infinite-dimensional group of diffeomorphisms on M.

[2]Most of these problems are facets of the so-called problem of time in quantum gravity [7–9].

section, we will present what we believe to be the most appropriate framework for theories of MBI, namely, histories theory.

3 Histories theory

In this section, we present the logical reformulation of physical theories that is based on *histories*, to which we will refer as *histories theory*. A history is a time-ordered sequence of properties of a physical system. In histories theory, any physical system is described in terms of (i) logical propositions about histories of the system and of (ii) the probabilities associated to such propositions. In particular, the notion of dynamics is incorporated into the rule of probability assignment[3]. The emphasis on the logical and probabilistic aspects of physical theories makes histories theory particularly suitable for the description of MBI, because propositions and probabilities are ontologically neutral. They can be meaningfully defined also for mental processes.

The idea of translating physics into the language of logical propositions originates from von Neumann and Birkhoff [10]. The idea that dynamics can be incorporated into the probability assignment for histories (in both classical or quantum physics) is due to Wigner and collaborators [11]. The logical reformulation of quantum mechanics in terms of histories is a key achievement of the consistent/decoherent histories approach to quantum theory by Griffiths, Omnés, Gell-Mann and Hartle [12–15]. This reformulation is the basis of our presentation here, as it can easily be adapted to any physical theory by using the *temporal logic* axiomatization developed by Isham [16, 17]. The consistent incorporation of dynamics in histories theory and the analysis of the theory's temporal structure is due to Savvidou [18, 19]. For the relation of histories theory to stochastic processes, see, Ref. [20] and for the histories theory version of GR, see, Ref. [21].

[3]Physical theories are traditionally described in terms of kinematics, dynamics and initial conditions. Kinematics defines the physical variables and the symmetries of a system, and how the former relate to measurable quantities. Dynamics describes how physical variables evolve in time. Initial conditions are necessary in order to obtain unique predictions about particular physical systems. In histories theory, the notion of kinematics is incorporated into the definition of the space of history propositions, while the notions of dynamics and initial conditions are incorporated into the rule of probability assignment.

3.1 History propositions

Consider an elementary physical system: a point-like particle moving in a line. This system is described by propositions such as the following.

- κ_t = "at time t, the particle's position x takes values between $3m$ and $5m$" (in a given coordinate system);
- λ_t = "at time t, the particle's momentum p takes values between $5kgm/s$ and $6kgm/s$";
- μ_t = "at time t, the particle's energy E takes values between $5J$ and $100J$".

The propositions above refer to a single moment of time t. For a given particle, they may be true or they may be false. It turns out that all single-time propositions about one particle can be expressed solely in terms of the particle's position and momentum. Each proposition corresponds to a subset C of the *state space* Γ, i.e., a set that consists of points (x, p). The elements of Γ are called *microstates* and they provide the most precise description of the system at one moment of time.

We can also consider *history propositions* for the particle, i.e. propositions that refer to more than one moments of time. The following are examples.

- α = "at time t_1, the particle's position x takes values between $3m$ and $5m$, and at time t_2 the particle's momentum takes values between $5kgm/s$ and $6kgm/s$"
- β = "at all times t between t_1 and t_2, the particle's momentum p takes values between $5kgm/s$ and $6kgm/s$ "
- γ = "at some time t between t_1 and t_2, the particle is recorded by a detector located at $x = 0$"

We will denote the set of history propositions of a system by $\mathcal{V}$.

History propositions correspond to subsets C of the *history space* Π. In order to define the latter, we first identify a *time-set* $\mathcal{T}$ that consists of all time instants t, and it is equipped with an ordering relation $\leq$, with the physical interpretation of "earlier than".

The history space is defined as the space of all paths on Γ, where a path $\xi : \mathcal{T} \to \Gamma$ is a function that assigns one point $\xi(t) \in \Gamma$ to each time $t \in \mathcal{T}$. The points of Π are called

fine-grained histories. They define propositions that give the most precise description of the physical system at all times (fine-grained propositions). History propositions that are not fine-grained are called *coarse-grained*. For example, in a theory where the fine-grained histories refer to atomic motions, any history about properties of neurons is coarse-grained.

We can combine history propositions using logical operators such as AND, OR, IMPLIES, NOT, and so on. Given two history propositions α and β, we can always define the history propositions $\alpha \,\text{AND}\, \beta$, $\alpha \,\text{OR}\, \beta$, $\alpha \,\text{IMPLIES}\, \beta$, and $\text{NOT}\, \alpha$. We also define the impossible history proposition $\emptyset$ (the proposition that can never be true) and the trivial history proposition $\mathbb{1}$ (the proposition that can never be false). We call two history propositions α and β *disjoint*, if there is no way that they can both be both true, i.e., if $\alpha \,\text{AND}\, \beta = \emptyset$. Coarse-grained propositions are obtained by joining disjoint fine-grained propositions through the connective OR.

Obviously, single-time propositions are special cases of history propositions. Some multi-time history propositions can be constructed from single-time propositions using the connective AND THEN of temporal conjunction. For example, the propositions α, κ_t and λ_t defined earlier satisfy

$$\alpha = \kappa_{t_1} \,\text{AND THEN}\, \lambda_{t_2}, \qquad t_1 < t_2. \tag{1}$$

In systems described by classical physics, we can always find a set Π of fine-grained histories, so that any history proposition α corresponds to a subset $C(\alpha)$ of Π. Hence, we can express all logical operations set-theoretically. For example, $C(\alpha \,\text{AND}\, \beta) = C(\alpha) \cap C(\beta)$, $C(\text{NOT}\, \alpha) = \Pi - C(\alpha)$, and so on. Hence, the set of history propositions has the structure of a Boolean algebra.

History propositions in quantum systems are very different. They do not have a Boolean algebra structure. Furthermore, there is an infinity of different fine-grained sets of histories, and each set defines a different physical description of a physical system. The mathematical structure of the space of history propositions is significantly more complex. Its intricacies, while crucial from the perspective of quantum foundations, are peripheral to the aims of this paper. A brief description of quantum history propositions and of their mathematical structure is given in the Appendix A. The reader may consult Ref. [16] for a detailed analysis.

3.2 Probability assignment

The predictions of all physical theories are expressed in terms of probabilities assigned to history propositions. A *partial probability function* Prob(·) is a rule that assigns a probability Prob(α) to history propositions α that belong in a subset $\mathcal{W}$ of the set $\mathcal{V}$; $\mathcal{W}$ is typically closed under the logical operations mentioned earlier. The restriction of history propositions to a subset $\mathcal{W}$ is a typical quantum phenomenon. Quantum probabilities are always defined in reference to a context, for example, a specific experimental configuration. If $\mathcal{W} = \mathcal{V}$, then we call Prob(·) a *complete probability function*.

Given a probability function Prob(·), one defines the *conditional probability* of a history proposition α given a history proposition β, as

$$\text{Prob}(\alpha|\beta) = \frac{\text{Prob}(\alpha \,\text{AND}\, \beta)}{\text{Prob}(\beta)}. \tag{2}$$

The physical predictions of the theory are typically expressed in terms of conditional probabilities Prob($\alpha|\beta$), where α refers to measurement outcomes.

Three distinct types of probability functions are used in physics, whereupon one talks about three types of *processes*: deterministic, stochastic and quantum. (Further types of processes are mathematically possible, but have not yet found use in physics.)

Deterministic processes have the following property. For *any* single-time proposition α_t, and a time $t' > t$, there is a unique minimal single-time proposition $\beta_{t'}$, such that Prob($\beta_{t'}|\alpha_t$) = 1. The proposition $\beta_{t'}$ is minimal in the sense that, if some other proposition $\gamma_{t'}$ satisfies Prob($\gamma_{t'}|\alpha_t$) = 1, then $\beta_{t'}$ IMPLIES $\gamma_{t'}$.

This definition of deterministic processes is restricted, as it ignores process with memory, but it suffices for present purposes. It characterizes all processes described by dynamical systems, i.e., by differential equations on the state space Γ, like Newton's equations of classical mechanics. The key point is that in deterministic processes, probabilities refer solely to the ignorance of the system's precise initial conditions.

Stochastic processes are characterised by a complete probability function on $\mathcal{V}$ that satisfies the Kolmogorov additivity condition: for all disjoint history propositions α and β,

$$\text{Prob}(\alpha \,\text{OR}\, \beta) = \text{Prob}(\alpha) + \text{Prob}(\beta). \tag{3}$$

Examples of stochastic processes are Brownian motion, random walks and evolutionary

processes. Strictly speaking, deterministic processes are a special case of stochastic processes.

Quantum processes are different. The natural probability function for quantum history propositions does not satisfy Kolmogorov's additivity condition for arbitrary history propositions α and β. This implies that in any given physical situation, probabilities cannot be assigned to all possible history propositions, but only to a specific subset thereof. Hence, quantum probability measures are partial. The exact specification of propositions to which probabilities can be assigned is a open issue in quantum foundations that is closely related to the quantum measurement problem. However, an uncontroversial choice is to restrict to history propositions that describe measurement outcomes in specific experiments. In this case, the probabilities associated to quantum processes coincides with the standard formulations of quantum theory in the Copenhagen interpretation [11, 20].

The three types of processes above are related. Deterministic processes can arise as limiting behavior of either stochastic or quantum processes, and stochastic processes can arise as limiting behavior of either deterministic or quantum processes. In what follows, we shall refer to deterministic and stochastic processes as *classical* processes, in the sense that they are compatible with classical physics (as contrasting quantum physics).

4 Histories theory description of MBI

4.1 Propositions about mental processes

In this section, we consider history propositions associated to a psycho-physical system and we argue that the most natural mathematical description of such systems introduces irreducibly mental degrees of freedom in addition to physical ones.

Consider a system that consists of Mary, a human person that lives in a closed room, together with all other physical objects in the room. We denote by $\mathcal{V}_\Phi$ the set of all history propositions about physical properties of the system. For example, $\mathcal{V}_\Phi$ contains propositions about a grey couch in the room, about Mary's movements as she sits on the couch, or about Mary's neurons firing while she sleeps on the couch. In principle, $\mathcal{V}_\Phi$ is fully determined from existing theories of physics.

There is also a set $\mathcal{C}$ of history propositions about mental properties in the system. Of

course, these properties refer to Mary and not to any other object in the room. $\mathcal{C}$ includes propositions about Mary's emotions, thoughts and qualia. Physicalist theories of mind would identify $\mathcal{C}$ with a subset of $\mathcal{V}_\Phi$. We will argue that it is more reasonable to assume that $\mathcal{C}$ is a subset of a different set $\mathcal{V}_\Psi$ of mental propositions that does not overlap with $\mathcal{V}_\Phi$. With this assumption, the set of all possible propositions about the system is the Cartesian product $\mathcal{V}_\Phi \times \mathcal{V}_\Psi$.

To this end, let us assume that Mary's room initially contains no green or red object. At time t_0 an object is inserted in the room. This object may be either a green pepper or a red rose. The pepper is green in the sense that it reflects light with wavelength of 500-550 nm; the rose is red in the sense that it reflects light with wavelength of 650-700 nm. Consider the history propositions

α_g = "A green pepper is inserted in the room at time t_0, and then Mary sees it",
α_r = "A red rose is inserted in the room at time t_0, and then Mary sees it".

By "Mary sees it" we mean a conjunction of propositions that include light from the object reaching Mary's retina, and an electrochemical signal carrying this particular information into the brain.

Next, we consider two history propositions that refer to mental properties,

β_G = "Mary has a GREEN experience at some time t after t_0",
β_R = "Mary has a RED experience at some time t after t_0".

GREEN and RED in capital letters refer to color qualia, i.e., individual instances of color experience [22]. We can avoid using the word "Mary" (which might require explaining what a person is) by rephrasing β_G as "There is a GREEN experience at some time t after t_0" and similarly for β_R.

Obviously, there is a strong correlation between α_g and β_G and between α_r and β_R. Consider an experiment in which either the rose or the pepper is inserted into the room and Mary telling us the color she sees. Repeating this experiment many times, we expect to find the probabilities,

$$\begin{aligned} \text{Prob}(\alpha_r \,\text{AND}\, \beta_R) = 1, \quad & \text{Prob}(\alpha_r \,\text{AND}\, \beta_G) = 0, \\ \text{Prob}(\alpha_g \,\text{AND}\, \beta_R) = 0, \quad & \text{Prob}(\alpha_g \,\text{AND}\, \beta_G) = 1, \end{aligned} \tag{4}$$

modulo some errors of order $\epsilon << 1$.

The aim of many physicalist research programs is to map β_R and β_G to elements of $\mathcal{V}_\Phi$ and to compute the probabilities (4) solely in terms of physics. The problem with this program is that existing physical theories are expressed in terms of particle properties, field properties, spacetime properties; qualia do not fit in.

Suppose then one finds a map that expresses β_R in terms of physical properties, i.e., that β_R logically coincides some proposition $\gamma_R \in \mathcal{V}_\Phi$. In what sense does the proposition γ_R depend on the quale RED? Since qualia appear neither in the construction of the space of propositions, nor in the probability assignment, RED can only be used as a label, i.e., as a non-dynamical index that identifies this proposition. It is certainly not a property to which the proposition refer. However, labels are arbitrary in physics: they are chosen as a matter of convention and they can be interchanged at will. This implies that there is no explanation from physics why one particular physical proposition $\alpha \in \mathcal{V}_\Phi$ is correlated to α_R and not to α_G, i.e., why a red rose corresponds to the experience RED and not to the experience GREEN[4].

Let us consider the situation formally, and ignore for the moment the meaning of the propositions. We have two sets of propositions $A = \{\alpha_r, \alpha_g\}$ and $B = \{\beta_R, \beta_G\}$,

(i) with strong probabilistic correlations given by Eq. (4);

(ii) with no known way of logically identifying elements of A with elements of B;

(iii) with strong arguments that such an identification may not be possible—see, [22] and references therein.

A physicist encountering this state of affairs in some problem would not hesitate to conclude that the correlations are dynamical. He or she would propose a model in which A

[4]This follows from an translation of Locke's famous argument about an 'inverted spectrum' [23]—see also Ref. [24] and references therein—into the language of contemporary physics. Let $i : \mathcal{C}_q \to \mathcal{V}_\Phi$ be the inclusion map of the set $\mathcal{C}_q$ of propositions about qualia into the set $\mathcal{V}_\Phi$ of history propositions about physical properties. Since qualia are not part of the physical theory, the physical predictions ought to be the same for any inclusion map $i' = f \circ i : \mathcal{C}_q \to \mathcal{V}_\Phi$, where f is an automorphism of $\mathcal{C}_q$. Thus, the probability assignment is invariant under the group $\mathrm{Aut}(\mathcal{C}_q)$ of automorphisms of $\mathcal{C}_q$. This means that there is no dynamical reason why the insertion of a red rose is correlated with a RED quale rather than a GREEN quale. As a matter of fact, there is no reason why the insertion of the rose is not correlated with a sound quale, i.e., why Mary does not hear Beethoven's 9th symphony on seeing the red rose.

describes one particular set of degrees of freedom and B describes a set of different degrees of freedom. These degrees of freedom interact dynamically in order to produce the probabilistic correlations of Eq. (4). The properties of the different degrees of freedom and the details of the interaction depend on the system under consideration, but the logic of the explanation is system-independent.

Suppose we transfer this way of thinking to the mind-body problem. We should introduce a set $\mathcal{V}_\Psi$ of history propositions about mental degrees of freedom, such that $B \subset \mathcal{V}_\Psi$. $\mathcal{V}_\Psi$ must be disjoint from $\mathcal{V}_\Phi$, i.e., $\mathcal{V}_\Psi$ and $\mathcal{V}_\Phi$ contain different propositions. Since $A \subset \mathcal{V}_\Phi$, the dynamical correlations of Eq. (4) are to be explained in terms of a probability rule for elements of the set $\mathcal{V}_\Phi \times \mathcal{V}_\Psi$.

The history propositions in $\mathcal{V}_\Psi$ and $\mathcal{V}_\Phi$ may be different, but the two sets must have isomorphic time-sets $\mathcal{T}_\Psi$ and $\mathcal{T}_\Phi$. This means that there is a bijection $f : \mathcal{T}_\Phi \to \mathcal{T}_\Psi$, such that $f(t_1) \leq f(t_2)$ for all $t_1 \leq t_2$. This is physically obvious: if Mary experiences GREEN before RED, then the time of the GREEN experience (as measured by a physical clock) must be prior to the time of the RED experience. In other words, psychological time and physical time have the same ordering.

4.2 The ΨΦI formalism

The arguments above suggest the following framework for theories with psycho-physical interaction (*ΨΦI formalism*).

1. *Causal structure.* The studied systems involve both physical and mental processes. All possible scenarios about a specific system can be expressed in terms of history propositions defined with respect to a time-set $\mathcal{T}$.

2. *Set of propositions.* History propositions have one physical and one mental component, i.e., they belong to the set $\mathcal{V}_\Phi \times \mathcal{V}_\Psi$, where $\mathcal{V}_\Phi$ contains history propositions about physical degrees of freedom and $\mathcal{V}_\Psi$ contains history propositions about mental degrees of freedom.

Purely physical propositions are of the form $(\alpha, \mathbb{1}_\Psi)$ and purely mental propositions are of the form $(\mathbb{1}_\Phi, \alpha)$. We will denote such propositions as α_Φ and α_Ψ, respectively. Hence, we can express a general history proposition $\alpha = \alpha_\Phi \text{ AND } \alpha_\Psi$, i.e., as a logical conjunction of its physical component α_Φ and its mental component α_Φ.

3. *Probability rule.* There is a class of partial probability functions $\mathrm{Prob}(\cdot)$ on $\mathcal{V}_\Phi \times \mathcal{V}_\Psi$ that determine physical predictions. The assumption of partial probability functions allows us to accommodate a quantum theory, without committing to a particular interpretation. The complexity of the quantum probability rule for histories implies that there are many more possible ways of expressing mental-physical interaction than in classical physics.

If we treat matter as classical, then we can take $\mathrm{Prob}(\cdot)$ to be a complete probability function.

The probability function incorporates non-trivial dynamical interaction between mental and physical degrees of freedom. This means that conditional probabilities of the form $\mathrm{Prob}(\alpha_\Phi|\beta_\Psi)$ and $\mathrm{Prob}(\gamma_\Psi|\delta_\Phi)$ have non-trivial dependency on propositions β_Ψ and δ_Φ, respectively.

4. *Limiting behavior.* Let us denote by Ω_Φ the proposition that there are no physical processes at any $t \in \mathcal{T}$. Ω_Φ is not to be confused with the impossible proposition $\emptyset_\Phi$. We also denote by Ω_Ψ the proposition that there are no mental processes at any $t \in \mathcal{T}$. We expect that in absence of mental processes, the probability function reduces to the known one of physics, denoted by Prob_Φ, i.e.,

$$\mathrm{Prob}(\alpha_\Phi \text{ AND } \Omega_\Psi) = \mathrm{Prob}_\Phi(\alpha_\Phi). \tag{5}$$

Unless we want to entertain the possibility of ghosts, we must postulate that no mental processes are possible in absence of physical processes, i.e.,

$$\mathrm{Prob}(\Omega_\Phi \text{ AND } \alpha_\Psi) = 0. \tag{6}$$

The principles above can be naturally incorporated into the temporal logic description of histories [16], and, thereby, provide an axiomatic characterization of the $\Psi\Phi$I formalism. Technical details in the formulation and elaboration of the axioms will be presented in a different publication. The key point is that the principles above define a general framework that can accommodate many different theories of MBI. Such theories will differ on the mathematical characterization of $\mathcal{V}_\Psi$ (the fundamental mental variables), and on the explicit construction of the probability function $\mathrm{Prob}(\cdot)$ (dynamics).

At the moment, we know the structure of $\mathcal{V}_\Phi$ and its associated probability rule. $\mathcal{V}_\Phi$ consists of all history propositions allowed by the Standard Model of particle physics, and

Prob(·) is the standard probability assignment for history propositions in quantum theory. Physics at the nuclear scale and beyond is most likely irrelevant to a mind-body coupling, so we can coarse-grain away Standard Model physics, and take $\mathcal{V}_\Phi$ to contain propositions about nuclei, electrons and the EM field; the probability assignment will again be quantum. It is a widely held belief among neuroscientists that we can coarse-grain even further, and that it suffices to work with a set $\mathcal{V}_\Phi$ of propositions about macromolecules or cells, subject to a *classical* probability assignment. In any case, the Φ part of a $\Psi\Phi$I theory is based on known physics.

In contrast, the Ψ part is largely unknown. We know many history propositions about mental processes at a phenomenological level, and these are elements of $\mathcal{V}_\Psi$. However, we know nothing about the mathematical structure of $\mathcal{V}_\Psi$, i.e., its fine-grained histories and how they can be joined through the OR operation in order to form coarse-grained propositions. Hence, the known elements of $\mathcal{V}_\Psi$ are disconnected from any underlying structure. The latter can be provided only by a mature mathematical theory of mind. It is quite possible that our familiar mental processes correspond to a highly coarse-grained mental history propositions, the same way that our everyday experience lies at a level of description much coarser than that of fundamental physics.

The probability assignment on $\mathcal{V}_\Phi \times \mathcal{V}_\Psi$ is also unknown, except for the condition that it reduces to the probability rules of physics in absence of mental phenomena. Since matter is fundamentally quantum mechanical, a fundamental *$\Psi\Phi$I theory cannot be classical*. Then, there are only three conceivable scenarios. Fundamental $\Psi\Phi$I processes are either (i) fully quantum, which means that mental processes should also be subject to the probabilistic rules of quantum mechanics, or (ii) they involve a quantum/classical hybrid (see, Sec. 5.3), or (iii) they follow a probability rule with no analogue in current physics.

The conclusion above does not preclude the use of classical $\Psi\Phi$I theories at coarser levels of description, since classical processes often emerge as limiting cases of quantum ones. Indeed, most neuroscience research proceeds under the assumption that the physical phenomena correlated to mental processes are essentially classical. This means that the relevant physical objects have lost their irreducible quantum features (except for the ones pertaining to the chemical properties of atoms). There are theoretical objections to this point of view, which originate from proposals that quantum phenomena are important for understanding consciousness, for example, Refs. [25–27]. We are agnostic on this issue. The

	General Relativity	**ΨΦI formalism**
Mathematical Framework (Dyn)	Lagrangian Field Theory	Histories Theory
Component A	spacetime geometry	mind
Representation of A	Lorentzian metric	fine-grained elements of $\mathcal{V}_\Psi$
Component B	matter	matter
Representation of B	matter fields	fine-grained elements of $\mathcal{V}_\Phi$
Interaction	Lagrangian dynamics	probabilities on $\mathcal{V}_\Phi \times \mathcal{V}_\Psi$

Table 1: Structural correspondence between the ΨΦI formalism and GR.

ΨΦI formalism works either way, but it is much easier to work with if quantum phenomena can be ignored.

In principle, classical probabilistic models for a few discrete degrees of freedom can be constructed directly from the ΨΦI axioms, without any knowledge of the deep structure of $\mathcal{V}_\Psi$. Such models could describe elementary mental processes, like, for example, the distinction of a small number of colors by a living person. It is too early to tell whether they could lead to testable predictions or not.

Finally, we remind the reader that the overall rationale of the ΨΦI formalism conforms to the strategy that was sketched in Sec. 2—see, Table (1) for the analogy to GR.

5 ΨΦI theories and physics

In the previous section, we argued that the histories description of physical theories can easily be extended to incorporate mental degrees of freedom. Here, we explore plausible properties of such theories, especially, in relation to open issues in physics.

5.1 Energy conservation

A popular objection to theories of MBI is that MBI conflicts with fundamental laws of physics, in particular the conservation of energy. A problem with this objection is that it assumes a 19th century understanding of physics. Today, we know that energy conservation

is not a universal law, as it does not hold in General Relativity[5] and it holds only with qualifications in quantum theory[6]. Thus, an MBI theory with no strict energy conservation is not problematic in an epistemic sense—for further discussion, see [28, 29].

More importantly, MBI does not always lead to a violation of energy conservation. In the Appendix, we study the status of energy conservation for classical ΨΦI processes. We identify dynamics that are fully compatible with energy conservation, in the sense that the MBI does not add or subtract energy to the physical degrees of freedom. Energy conservation is a consequence of a particular *coupling* between the mental and the physical degrees of freedom. This coupling is neither artificial nor contrived: it is mathematically elegant and simple, as befits a fundamental theory. It can also be generalized for quantum systems.

Energy-conserving ΨΦI theories are aesthetically appealing, but certainly, other options are available. For example, one may consider MBI dynamics that conserve a generalised notion of energy. In mechanical systems, energy conservation is a *consequence* of the symmetry of time translation, i.e., the requirement that the dynamics is unchanged under a transformation that moves the time of the various events by a constant amount. If a ΨΦI theory shares this symmetry, then a *generalised energy* variable—that depends on both physical and mental degrees of freedom—is plausibly conserved. Hence, we can continue to use our current notion of energy, provided we include a contribution from mental processes. After all, the notion of energy has been generalised many times ever since its inception, as it has been applied to increasingly broader categories of phenomena.

It is also possible that a ΨΦI theory does not admit either energy conservation or

[5]Energy is conserved only in stationary spacetimes, i.e., a class of spacetimes characterized by a specific symmetry (the existence of a timelike Killing field). Energy is *not* conserved in generic spacetimes. For example, energy is not conserved in the expanding universe models that are employed in cosmology. Energy conservation holds approximately at scales much smaller than the Hubble length that characterises the expansion of the universe.

[6]In quantum theory, energy is strictly conserved only for a particular class of initial states (namely, the eigenstates of the Hamiltonian operator). In general, what is conserved (i.e., it is the same at all times) is the probability distribution for the values of energy. This means that two individual quantum systems that have been prepared identically will, in general, be measured with different values of energy. Of course, a measured system is an open system, so one does not expect energy to be preserved during the measurement process. In any case, the statement that the value of energy remains the same during time evolution is not true.

conservation of some generalised notion of energy. In this case, energy conservation is an approximate conservation law that holds in regimes where the MBI is negligible. Approximate conservation laws are quite common in physics, they apply to physical quantities that are conserved in a large classes of interactions but not in all. One example is isospin which is conserved in strong and electromagnetic interactions but not in weak interactions.

5.2 Threshold states

We believe that a living person has mental activity, but we ordinarily deny such activity to the dead body of the same person, as we deny it to rocks, cars, electrons or computers. Thus, we believe that in some systems mental processes never occur, and that in some other systems mental processes sometimes occur and sometimes do not. In a ΨΦI theory, we implement such distinctions by introducing a set of propositions $\Omega_t \in \mathcal{V}_\Psi$ which assert that no mental processes take place at time t. For rocks, cars and so on, the probabilities associated to propositions of $\mathcal{V}_\Psi$ other than Ω_t are always zero. For living people, probabilities for propositions other than Ω_t can obviously be non-zero.

The analogue of Ω_t in physics is the *vacuum* of quantum field theory (QFT). The vacuum is the state of the system in which no particles (of a given type A) are present. In QFT, one is often interested in the generation of A particles from the vacuum, for example, in the presence of other particles or external fields. These phenomena reveal the interaction channels of the A-particles with the rest of the world.

We can follow an analogous reasoning in the ΨΦI formalism. Let C denote a configuration of a physical system, i.e., a subset of the system's state space Γ. We represent the proposition that C is present at time t by C_t. The family of history propositions

$$\eta(C) = (\Omega_t \,\text{AND}\, C_t) \,\text{AND THEN}\, (\text{NOT}\, \Omega_{t'}) \tag{7}$$

for $t' > t$ describes the generation of mental states out of the mental 'vacuum' Ω. If C describes a rock, a car, or a dead body, any reasonable probability assignment will give $\text{Prob}[\eta(C)] = 0$. The same holds if C describes a living person. However, there exist physical configurations C for which $\text{Prob}[\eta(C)] > 0$. Such configurations are responsible for the generation of (proto)mental states from non-mental states: they are the *threshold* to the world of mental processes—see, Fig. 1.

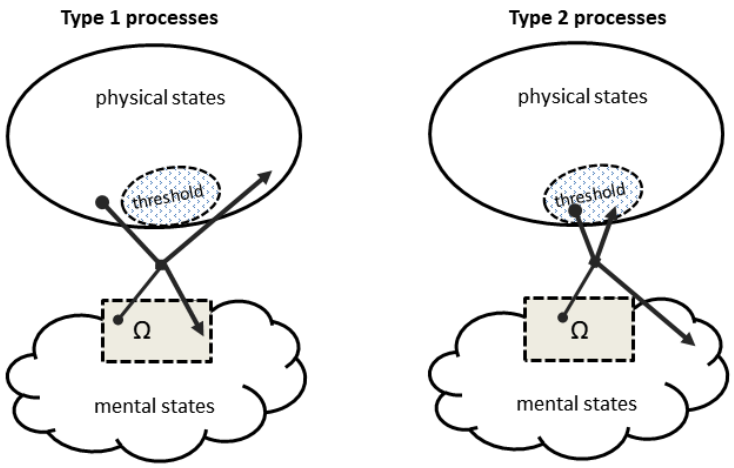

Figure 1: A graphical description of threshold states. We assume that mental degrees of freedom are initially in their 'vacuum' Ω. Processes of type 1 cannot generate mental states out of Ω. Only if the physical state is a threshold state (processes of type 2) is it possible to reach a non-trivial mental state.

A formal definition of threshold states should also involve some condition of minimality, otherwise the whole Earth 4.5 billion years ago would qualify as a threshold state. Hence, we require that a threshold state satisfies $\text{Prob}[\eta(C)] > 0$ and also $\text{Prob}[\eta(D)] = 0$, for any $D \subset C$.

Threshold states are crucial for understanding how organisms with mental processes emerged in the history of life. Furthermore, if the analogy to QFT is valid, the properties of threshold configurations may suggest the form of the MBI-generating terms in the probability assignment.

In QFT, the defining feature of threshold configurations is energy: no A particle can be generated from the vacuum unless the available energy is greater than the rest mass m_A of A. When asking what is the analogue of energy for the threshold configurations of a $\Psi\Phi$I theory, there seems to be an obvious answer: *information*. Many mental processes can be described in terms of information processing, and, of course, the brain is an information-processing system. Like causal ordering, the notion of information strands both sides of the mental-physical divide. Indeed, it has been proposed as a crucial component of psychophysical theories [3]. It is therefore plausible that threshold states are characterized

by high informational capacity.

The problem here is that information theory concepts are not fundamental to our current physical theories. They only provide an additional layer of interpretation and some technical tools. Notions such as informational capacity or information processing are not absolute properties of physical systems. One deterministic process is as good as another in terms of information processing: there is no criterion for distinguishing the electric currents inside a laptop from the motions of air molecules in an empty bottle. We call the former information processing because of their meaningful output: the electric currents cause a complex production of light on the liquid crystal display that we interpret as text. This criterion presupposes us, i.e., the existence of mindful observers.

It is a plausible conjecture that the notions of information processing and informational capacity are fundamentally defined in terms of mental processes, and not physical ones. This would imply that threshold states 'distribute' the mental concept of information to the physical degrees of freedom. Hence, MBI could lead to a fundamental definition of information in physics.

5.3 MBI and quantum state reduction

The *measurement problem* in quantum theory is that quantum theory does not explain the emergence of definite properties for physical systems (for example, measurement outcomes). Definite properties are not, in general, compatible with quantum processes. One possible resolution, suggested by the founders of quantum theory, is the use of *irreducibly classical concepts* for the description of the measurement device. But what constitutes a measuring device? If the a particle's position is correlated to the reading of a pointer in an apparatus, should we describe the pointer classically? What if the pointer is recorded by a camera? Should we treat the pointer as a quantum system and the camera as classical? The boundary of the quantum/classical divide appears arbitrary.

The problem is aggravated in interpretations of quantum theory that treat the quantum state as an objective feature of a physical system. In these interpretations, the change of state after measurement (quantum state reduction) is a physical process. Then, the arbitrariness of the quantum/classical split implies an ambiguity in the physical description. Von Neumann and Wigner proposed that the quantum/classical split and the body/mind split coincide: the quantum state is reduced when the result of a measurement enters the

observer's consciousness [30, 31].

In contrast, *dynamical state reduction* models [32] postulate that there is a tiny probability of reduction for each particle's wavefunction, which adds up to be significant for systems that contain a huge number of particles. This explains why macroscopic measuring apparatuses behave classically. Alternatively, one may postulate irreducibly classical entities that universally interact with quantum systems. Barring consciousness, the main candidate for this role is the gravitational field [33–37]. The only way to consistently formulate a quantum-classical interaction involves dynamical state reduction for the quantum system; again, this process can cause measurements apparatuses to always behave classically.

The latter scenario is relevant to ΨΦI theories. If mental processes are treated classically, the resulting ΨΦI theory involves coupling of quantum to classical variables. Hence, quantum systems undergo dynamical state reduction as a result of MBI. In other words, dynamical reduction is a natural candidate for the physical channel through which mind acts upon matter.

We do *not* propose ΨΦI theories as a solution to the quantum measurement problem. We think that consciousness-based solutions to the measurement problem are highly counterintuitive in the context of quantum cosmology. They seem to imply that no definite properties could exist prior to the emergence of the first conscious observers. Nonetheless, the ΨΦI formalism can, in principle, be used in order to construct predictive models of the von Neumann-Wigner idea.

Reduction in ΨΦI theories is not the universally occurring process postulated by dynamical reduction models. Most probably, it would only occur in the nervous system of biological organisms. Interestingly, ΨΦI predictions may turn out to be compatible with proposals that relate dynamical reduction in the brain to consciousness—like, for example, the theory of Orchestrated Objective Reduction by Hameroff and Penrose [26]—even if the latter treats the physical world as causally closed.

5.4 Maxwell's demon

The *Maxwell-demon* paradox is a thought experiment proposed by J. C. Maxwell in 1867, according to which a demon can violate the second law of thermodynamics [38]. The demon is a being that controls a small door between two chambers of gas. As individual

gas molecules approach the door, the demon quickly opens and shuts the door so that fast molecules pass into the other chamber, while slow molecules remain in the first chamber. The demon's action causes one chamber to warm up and the other to cool, thus decreasing entropy and violating the Second Law of Thermodynamics.

An immense number of papers have been written about Maxwell's demon, mainly aiming to remove the challenge to the Second Law. The key idea is that the process of information acquisition has an entropy cost, so that the total entropy is not reduced when the demon operates. While this research has led to many insights about the relation between information and statistical mechanics, the demon has not been permanently exorcised:

In so far as the Demon is a thermodynamic system already governed by the Second Law, no further supposition about information and entropy is needed to save the Second Law. In so far as the Demon fails to be such a system, no supposition about the entropy cost of information acquisition and processing can save the Second Law from the Demon [39].

In accordance with Maxwell's original conception, theories with MBI severely threaten the Second Law. To exorcise the Demon, we must restrict to ΨΦI models that are compatible with the laws of thermodynamics, or possibly with a reasonable generalization thereof. This generalization may involve, for example, a redefinition of entropy that incorporates acquisition of information by mindful agents.

6 Conclusions

The main aim of this paper is to show that the popular assertion that MBI is incompatible with physics is wrong. This was achieved by the construction of a mathematical framework that enables the construction of theories with MBI as an extension of current physical theories. The ΨΦI formalism originates from the histories formulation of physical theories, and it describes irreducibly mental degrees of freedom that interact with the physical degrees of freedom.

The ΨΦI formalism can incorporate any mental concept that can be expressed in terms of abstract structural and causal relations. Certainly, there are aspects of the mind that go beyond such relations; they cannot be described by the formalism. This is not a problem. GR demonstrates that even a fundamental theory does not require a faithful representation of the entities it describes, in order to be highly successful.

Many mental processes involve conscious experience. We can represent mathematically a conscious experience, by introducing a variable, say Con, that takes value 1 on all states that involve conscious experience and 0 otherwise. The time evolution and causal properties of the variable Con can then be studied for any psycho-physical configuration. Obviously, the $\Psi\Phi$I formalism cannot explain what "conscious experience" is. The existence of conscious experience has to be taken as a *brute fact* about the mind that defines the fundamental building blocks of a theory.

Physicalist theories of mind often invoke the remarkable success of physics to explain a huge number of phenomena through reduction to elementary physical processes. In our opinion, such arguments only pay lip-service to physics, while ignoring its history and actual research practice. The point is that neither reduction nor supervenience have been particularly successful as research strategies in physics.

The key step in constructing *fundamental* physical theories has always been the identification of the appropriate degrees of freedom for the problem at hand. In all major discoveries, it was necessary to introduce new degrees of freedom, well beyond the ones that were known at the time [40]. Examples include the electromagnetic field, the Rutherford-Bohr model of the atom, the concept of particle spin, the spacetime metric in GR, and the large number of new particles, charges and fields that had to be postulated in order to construct the Standard Model of strong and electroweak interactions.

Ever since Newton, the most successful strategy in physics has been the search for theories that unify seemingly very different phenomena by incorporating them in an overarching mathematical structure. Hence, the $\Psi\Phi$I formalism is much closer to the research practices of physics than any physicalist research program.

References

[1] Dirac, P. (1955), public lecture in the Indian Science Congress, Baroda.

[2] Chalmers, D. (1995), Facing Up to the Problem of Consciousness, *J. Consc. Stud.* 2: 200.

[3] Chalmers, D. (1995), The Puzzle of Conscious Experience, *Sci. Am.* 273, 80.

[4] Goldstein, H., Poole, C. P. and Safko, J.L. (2001), chapter 13 in *Classical Mechanics* (third edition), Reading: Addison Wesley.

[5] Harman P.M. (1982), *Energy, Force and Matter: The Conceptual Development of Nineteenth-Century Physics*, Cambridge: Cambridge University Press.

[6] Wheeler, J.A. and Ford, K. (1998), *Geons, Black Holes and Quantum Foam*, New York: W. W. Norton.

[7] Isham C. J. (1992), Canonical Quantum Gravity and the Problem of Time, Lecture Notes at Salamanca Summer School, gr-qc/9210011.

[8] Kuchar K. (1991),The Problem of Time in Canonical Quantum Gravity, in *Conceptual Problems of Quantum Gravity*, edited by A. Ashtekar and J. Stachel, Basel: Birkhäuser.

[9] Anderson E. (2017), *The Problem of Time: Quantum Mechanics Versus General Relativity*, Berlin: Springer.

[10] Birkhoff, G. and von Neumann J. (1936), The Logic of Quantum Mechanics, *Ann. Math.* 37: 823.

[11] Houtappel, R. M. F., van Dam, H. and Wigner, E. P. (1965), The Conceptual Basis and Use of the Geometric Invariance Principles, *Rev. Mod. Phys.* 37, 595.

[12] Griffiths, R. B. (2003), *Consistent Quantum Theory*, Cambridge: Cambridge University Press.

[13] Omnés, R. (1994), *The Interpretation of Quantum Mechanics*, Princeton: Princeton University Press; R. Omnés, (1999), *Understanding Quantum Mechanics*, Princeton: Princeton University Press.

[14] Gell-Mann, M. and Hartle, J. B. (1990), Quantum Mechanics in the Light of Quantum Cosmology, in *Complexity, Entropy, and the Physics of Information* ed. by W. Zurek, Reading: Addison Wesley.

[15] Hartle, J. B. (1995), Spacetime Quantum Mechanics and the Quantum Mechanics of Spacetime in *Gravitation and Quantizations*, Proceedings of the 1992 Les Houches Summer School, ed. by B. Julia and J. Zinn- Justin, Les Houches Summer School Proceedings, Vol. LVII, Amsterdam: North Holland. [gr-qc/9304006].

[16] Isham, C. J. (1994), Quantum Logic and the Histories Approach to Quantum Theory, *J. Math. Phys.* 35: 2157.

[17] Isham, C. J. and Linden, N. (1994), Quantum Temporal Logic and Decoherence Functionals in the Histories Approach to Generalised Quantum Theory, *J. Math. Phys.* 35: 5452.

[18] Savvidou, K. (1999), The Action Operator for Continuous-time Histories, *J. Math. Phys.* 40: 5657.

[19] Savvidou N. (2009), Space-time Symmetries in Histories Canonical Gravity, in *Approaches to Quantum Gravity*, edited by D. Oriti, Cambridge: Cambridge University Press.

[20] Anastopoulos, C. (2003), Quantum Processes on Phase Space, *Ann. Phys.* 303: 275.

[21] Savvidou, K. (2004), General Relativity Histories Theory I: The Spacetime Character of the Canonical Description, *Class. Quant. Grav.* 21: 615; General Relativity Histories Theory II: Invariance Groups, *Class.Quant.Grav.* 21: 631.

[22] Tye, M. (2016), Qualia, *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.),
URL = https://plato.stanford.edu/archives/win2016/entries/qualia/.

[23] Locke, J. (1689/1975), *Essay Concerning Human Understanding*, Oxford: Oxford University Press.

[24] Byrne, A. (2016), Inverted Qualia, *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.),
URL = https://plato.stanford.edu/archives/win2016/entries/qualia-inverted/.

[25] Beck, F., and Eccles, J. (1992), Quantum Aspects of Brain Activity and the Role of Consciousness, *Proc. Nat. Acad. Sci (USA)* 89: 11357.

[26] Hameroff, S. R. and Penrose R. (2014), Consciousness in the Universe: A Review of the 'Orch OR' Theory, *Phys. Life Rev.* 11: 39.

[27] Schwartz, J. M, Stapp, H. P. and Bauregard, M. (2005), Quantum Physics in Neuroscience and Psychology: a Neurophysical Model of Mind-Brain Interaction, *Philos. Trans. R. Soc. Lond. B* 360: 1309.

[28] Collins, R. (2008), Modern Physics and the Energy-Conservation Objection to Mind-Body Dualism, *American Philosophical Quarterly*, 45: 31.

[29] Pitts, J.B. (2019), Conservation Laws and the Philosophy of Mind: Opening the Black Box, Finding a Mirror, *Philosophia*. https://doi.org/10.1007/s11406-019-00102-7

[30] Von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics*, Princeton: Princeton University Press.

[31] Wigner, E. P. (1967), Remarks on the Mind Body Question, *Am. J. Phys.* 35: 1169.

[32] Bassi, A. and Ghirardi, G. C. (2003), Dynamical reduction models, *Phys. Rep.* 379: 257.

[33] Karolyhazy F. (1966), Gravitation and Quantum Mechanics of Macroscopic Objects, *Nuovo Cim.* 52:390.

[34] Diosi L. (1984), Gravitation and Quantum-Mechanical Localization of Macro-Objects, *Phys. Lett. A* 105: 199.

[35] Diosi L. (1987), A Universal Master Equation for the Gravitational Violation of Quantum Mechanics, *Phys. Lett.* 120: 377.

[36] Penrose R. (1986), Gravity and State Vector Reduction, in *Quantum Concepts in Space and Time*, R. Penrose and C. J. Isham editors, Oxford: Clarendon Press)

[37] Penrose R. (1996), On Gravity's Role in Quantum State Reduction, *Gen. Rel. Grav.* 28: 581.

[38] Maxwell, J. C. (1871), Chapter 12 of *Theory of Heat*; London: Longmans, Green, and Co.

[39] Earman, J. and Norton, J. D. (1999), EXORCIST XIV: The Wrath of Maxwell' s Demon. Part II. From Szilard to Landauer and Beyond, *Stud. Hist. Phil. Mod. Phys.* 30: 1.

[40] Anastopoulos, C. (2008), *Particle or Wave: The Evolution of the Concept of Matter in Modern Physics*, Princeton: Princeton University Press.

[41] E. G. Beltrametti and G. Cassinelli, *The Logic of Quantum Mechanics* (Cambridge University Press, 2010).

[42] Marsden, J.E. and Ratiu, T. (1999), *Introduction to Mechanics and Symmetry*, Berlin: Springer.

A The structure of histories theory

In this Appendix, we briefly summarize some features of histories theory, both classical and quantum.

A.1 Classical Histories Theory

Consider a system described by a sample space Γ that is a differentiable manifold. Single-time propositions corresponds to measurable subsets C of Γ. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ be a discrete time-set for this system with $t_1 < t_2 < \ldots < t_n$. In general, $\mathcal{T}$ may be any partially ordered set, but here we restrict our considerations to the simplest case.

The path space Π consists of all maps $\gamma : \mathcal{T} \to \Gamma$. The set of history propositions $\mathcal{V}$ consists of all measurable subsets of Π. Hence, we can write a history proposition for this system as

$$\alpha = (C_{t_1}, C_{t_2}, \ldots, C_{t_n}), \tag{8}$$

where C_{t_i} is a subset of Γ that corresponds to the proposition that the system was found in C_{t_i} at time t_i. Logical operations are standardly defined in terms of set-theoretic operations between subsets of Π.

A probability functional assigns a probability $\mathrm{Prob}(\alpha)$ to each history proposition $\alpha \in \mathcal{V}$, as

$$\mathrm{Prob}(\alpha) = \int dx_1 \ldots dx_n \chi_{C_{t_1}}(x_1)\chi_{C_{t_2}}(x_2)\ldots\chi_{C_{t_n}}(x_n)p(x_1,t_1;x_2,t_2;\ldots,x_n,t_n), \tag{9}$$

where $p(x_1,t_1;x_2,t_2;\ldots,x_n,t_n)$ is a probability measure on Π, and χ_C the characteristic function of the set $C \subset \Gamma$.

The theory of stochastic processes is an example of a classical history theory. Of particular interest are Markovian processes, which describe probabilistic systems without memory. The associate probability measures are of the form

$$p(x_1,t_1;x_2,t_2;\ldots,x_n,t_n) = \rho_{t_1}(x_1)g(x_2,t_2|x_1,t_1)g(x_3,t_3|x_2,t_2)\ldots g(x_n,t_n|x_{n-1},t_{n-1}), \tag{10}$$

where ρ_{t_1} is a probability density on Γ at the initial moment of time, and $g(x_2,t_2|x_1,t_1) > 0$ is the transition matrix between times t_1 and t_2. The transition matrix is stochastic, i.e., it satisfies $\int dx_2 g(x_2,t_2|x_1,t_1) = 1$.

Deterministic processes (without memory) correspond to the degenerate case where the transition matrix is a delta function, i.e.,

$$g(x_2, t_2|x_1, t_1) = \delta[x_2 - f_{t_2,t_1}(x_1)], \tag{11}$$

where $f_{t,t'}$ is a diffeomorphism on Γ indexed by t and t'. For classical mechanics and field theory, see, Ref. [19].

A.2 Quantum histories theory

In Copenhagen quantum mechanics, propositions correspond to measurement outcomes. Ideal measurements are described in terms of projection-valued measures (PVMs) on a Hilbert space $\mathcal{H}$, i.e., a family of projectors $\hat{P}_a$ indexed by a, that satisfy the following properties: (i) mutual exclusion, $\hat{P}_a\hat{P}_b = \delta_{ab}\hat{P}_a$, and (ii) exhaustion $\sum_a \hat{P}_a = \hat{I}$.

A sequence of measurements at times $t_1, t_2, \ldots, t_n$ corresponds to a sequence of PVMs $\hat{P}_{a_i,t_i}$ indexed by the time-parameter. A history of measurement outcomes is a sequence of projectors

$$\alpha := (\hat{P}_{a_1,t_1}, \hat{P}_{a_2,t_2}, \ldots, \hat{P}_{a_n,t_n}). \tag{12}$$

Suppose that the system is prepared at the state $\hat{\rho}_0$ at $t = 0$, and that its Hamiltonian is $\hat{H}$. The probability associated to α is

$$\text{Prob}(\alpha) = Tr\left(\hat{C}_\alpha\hat{\rho}_0\hat{C}_\alpha^\dagger\right), \tag{13}$$

where

$$\hat{C}_\alpha = \hat{P}_{a_n,t_n}(t_n)\ldots\hat{P}_{a_2,t_2}\hat{P}_{a_1,t_1}(t_1) \tag{14}$$

is an operator associated to the history α, defined in terms of the Heisenberg-picture projectors $\hat{P}_{a_i,t_i}(t_i) = e^{i\hat{H}t_i}\hat{P}_{a_i,t_i}e^{-i\hat{H}t_i}$.

The consistent/decoherent histories interpretation of quantum mechanics starts with the observation that the sequence $(\hat{P}_{a_1,t_1}, \hat{P}_{a_2,t_2}, \ldots, \hat{P}_{a_n,t_n})$ can be interpreted as referring to propositions about properties of a physical system, and not only to measurement outcomes. Then, it is possible to define logical operations, such as AND, OR, NOT and so on, between those history propositions, and to define a set $\mathcal{V}$ if history propositions that is closed under those operations.

The set of history propositions is constructed in terms of the history Hilbert space $\mathcal{K}_{his}$, defined as the tensor product of single time Hilbert spaces,

$$\mathcal{K}_{his} = \otimes_{t\in\mathcal{T}}\mathcal{H}_t, \tag{15}$$

where t is an element of the time-set $\mathcal{T}$, and $\mathcal{H}$ is a copy of the single-time Hilbert space indexed by t. The history (12) is represented by a projection operator $\hat{E} = \hat{P}_{a_1,t_1} \otimes \hat{P}_{a_2,t_2} \otimes \ldots \otimes \hat{P}_{a_n,t_n}$ on $\mathcal{K}_{his}$. General history propositions are represented by projectors on $\mathcal{K}_{his}$ that are not factorized. Then, the set $\mathcal{V}$ of history propositions coincides with the lattice $\Lambda(\mathcal{K}_{his})$ of projection operators on $\mathcal{K}_{his}$. Hence, the logical operations on $\mathcal{V}$ coincide with the lattice operations of $\Lambda(\mathcal{K}_{his})$ [41].

The incorporation of dynamics into the histories description is rather intricate, and it requires a detailed analysis of the symmetries of the formalism. For continuous time, dynamics are implemented through an action operator that is defined on $\mathcal{K}_{his}$ [18].

The rule (13) does not define a probability measure on the set of history propositions. Probabilities can only be defined with respect to a given *context*. In the Copenhagen interpretation, the context is given by the measurement set-up, i.e., the choice of the different PVMs at different moments of time.

In decoherent histories, the context is expressed in terms of the abstract concept of a *consistent set*. To this end, we define the decoherence functional for a pair of histories α and β of the form (13), as

$$d(\alpha,\beta) = Tr\left(\hat{C}_\alpha \hat{\rho}_0 \hat{C}_\beta^\dagger\right). \tag{16}$$

The decoherence functional is extended to general history propositions by the requirements of

- linearity: $d(\alpha \text{ OR } \gamma, \beta) = d(\alpha,\beta) + d(\gamma,\beta)$, for α AND $\gamma = \emptyset$, and
- hermiticity: $d(\alpha,\beta) = d(\beta,\alpha)^*$.

Then, d defines a bilinear map on $\mathcal{V}\times\mathcal{V}$.

Consider a set $\mathcal{W}$ of history propositions that are mutually exclusive and exhaustive. We call $\mathcal{W}$ a consistent set, if it satisfies the consistency condition

$$\text{Re } d(\alpha,\beta) = 0,\ \alpha \neq \beta, \tag{17}$$

for all $\alpha, \beta \in \mathcal{W}$. Then, the diagonal elements of the decoherence functional define a probability measure on $\mathcal{S}$,

$$\mathrm{Prob}_{\mathcal{W}}(\alpha) = d(\alpha, \alpha), \text{for all } \alpha \in \mathcal{W}. \tag{18}$$

Hence, in quantum theory we do not have a complete probability function on $\mathcal{V}$, but a family of partial probability functions $\mathrm{Prob}_{\mathcal{W}}$, each corresponding to a different consistent set $\mathcal{W}$.

B Energy conservation in ΨΦI theories

In this section, we explore the issue of energy conservation in ΨΦI theories. We consider classical processes, because energy conservation is precisely formulated in classical mechanics. However, some of the results can be appropriately generalised also for quantum systems.

B.1 Energy conserving dynamics

We first consider deterministic processes. They describe dynamical systems, i.e., systems with time evolution determined by a set of differential equations. These equations can be interpreted as a flow on a differentiable manifold that plays the role of the state space. For MBI models, the state space is of the form $\Gamma_\Phi \times \Gamma_\Psi$, where Γ_Φ contains the physical degrees of freedom and Γ_Ψ contains the mental degrees of freedom.

We describe the physical degrees of freedom in terms of classical mechanics, so that energy conservation is well-defined. Let us denote the points of Γ_Φ by ξ^a, for some discrete index a. Time evolution is given by Hamilton's equation

$$\dot{\xi}^a = \mathcal{P}^{ab}\partial_b H, \tag{19}$$

where H is the Hamiltonian, i.e., a function on Γ_Φ whose values correspond to the usual notion of energy.

The *Poisson tensor* $\mathcal{P}^{ab}$ in Eq. (19) is antisymmetric, non-degenerate and satisfies Jacobi's identity

$$\mathcal{P}^{ad}\partial_d\mathcal{P}^{bc} + \mathcal{P}^{cd}\partial_d\mathcal{P}^{ab} + \mathcal{P}^{bd}\partial_d\mathcal{P}^{ca} = 0. \tag{20}$$

The antisymmetry of $\mathcal{P}$ guarantees the conservation of energy, since

$$\dot{H} = \partial_a H \dot{\xi}^a = \mathcal{P}^{ab} \partial_a H \partial_b H = 0. \tag{21}$$

The most general dynamical system on $\Gamma_\Phi \times \Gamma_\Psi$, compatible with Eq. (19), is of the form

$$\dot{\xi}^a = \mathcal{P}^{ab}(\xi) \partial_b H(\xi) + G^a(\xi, y) \tag{22}$$

$$\dot{y}^i = F^i(y) + J^i(\xi, y), \tag{23}$$

where we denoted the points of Γ_Ψ by y^i; i is a discrete index that labels the mental degrees of freedom. The vector F^i on Γ_Ψ describes self-dynamics on Γ_Ψ. Interaction is described by the vector fields G^a on Γ_Φ and J^i on Γ_Ψ.

Eq. (22) implies that

$$\dot{H} = G^a \partial_a H. \tag{24}$$

Hence, energy is conserved if $G^a \partial_a H = 0$, i.e., if G^a is tangent to the surfaces of constant energy. To motivate our subsequent analysis, let us first assume that physical degrees of freedom satisfy some form of Hamilton equations, even in the presence of MBI. This implies that G^a is a Hamiltonian vector field for each y, i.e., that

$$G^a(\xi, y) = \mathcal{P}^{ab} \partial_b S(\xi, y) \tag{25}$$

for some scalar function S. Hence, energy is conserved if $\mathcal{P}^{ab} \partial_a S \partial_b H = 0$.

A Hamiltonian system with a large number of degrees of freedom may have constants of the motion other than the Hamiltonian. However, it is highly implausible that we can ever relate such constants, when they are defined for very different systems, for example, a room with one person inside and a concert hall with one thousand persons.

We expect that energy-conserving dynamics of sufficient generality are possible only if S depends on ξ solely through the Hamiltonian, i.,e., if $S(\xi, y) = \phi(H(\xi), y)$ for some function $\phi : \boldsymbol{R} \times \Gamma_\Psi \to \boldsymbol{R}$. Then, the evolution equation on Γ_Φ becomes

$$\dot{\xi}^a = \tilde{\mathcal{P}}^{ab}(\xi, y) \partial_b H(\xi) \tag{26}$$

where $\tilde{\mathcal{P}}^{ab}(\xi, y) = [1 + \phi'[H(\xi), y]] \mathcal{P}^{ab}(\xi)$, and the prime denotes the derivative of ϕ with respect to its first argument. Hence, energy is conserved if the mental degrees of freedom are coupled to the physical degrees of freedom through an antisymmetric tensor $\tilde{\mathcal{P}}^{ab}$.

The tensor $\tilde{\mathcal{P}}^{ab}$ is not an actual Poisson tensor, because it does not satisfy the Jacobi identity, Eq. (20). However, a small modification of our previous analysis can justify a Poisson tensor $\tilde{\mathcal{P}}^{ab}$ in Eq. (26), leading to a $\Psi\Phi$I theory with a much more interesting mathematical structure.

To this end, let us assume that there is a submanifold $\Omega \subset \Gamma_\Psi$ that corresponds to the mental vacuum of Sec. 5.2, i.e., if $y \in \Omega$ then no mental processes are present. Consider now a Poisson tensor $\tilde{\mathcal{P}}^{ab}$ that satisfies $\tilde{\mathcal{P}}^{ab} = \mathcal{P}^{ab}$, for any $y \in \Omega$. Then, Eqs. (26) and (23) define a dynamical system for MBI that reduces to the classical equations of motions in absence of mental processes. Hence, they define a deterministic $\Psi\Phi$I theory with energy conservation.

The difference from our previous analysis is that G^a is not a Hamiltonian vector field, i.e., Eq. (25) does not hold. Instead, $G^a = \Delta\mathcal{P}^{ab}\partial_b H$, where $\Delta\mathcal{P}^{ab} = \tilde{\mathcal{P}}^{ab} - \mathcal{P}^{ab}$.

The implementation of energy conservation constraints only one of the two equations that describe the dynamical system, namely, Eq. (26) Thus, it persists even if the term $F^i(y)$ of Eq. (23) is a 'random force', i.e., if the values of F^i at each moment of time are distributed probabilistically. Thus, energy conservation also applies to stochastic $\Psi\Phi$I theories—in fact it is completely insensitive to the dynamics of the mental degrees of freedom—, as long as Eq. (26) is satisfied.

The same structure can be employed in order to define energy-conserving dynamics for MBI theories, where matter is treated quantum mechanically. Schrödinger's equation on a Hilbert space $\mathcal{H}$ is equivalent to Hamilton's equation on the projective Hilbert space $\mathcal{PH}$ [42]. Hence, the above analysis can straightforwardly be transferred into a quantum context.

B.2 Generalised energy

In mechanical systems, energy conservation arises as a consequence of the time-translation symmetry, i.e., the fact that dynamics are unchanged under a transformation that moves the time of the various events. It is plausible that the time translation symmetry of the extended dynamics (22—23) corresponds to a conserved quantity $K(\xi, y)$ that reduces to the Hamiltonian H in a limit where the mind-body interaction term can be ignored. For example, $K(\xi, y)$ may be of the form $L(y) + H(\xi)$, where $L(y)$ is invariant under the self

dynamics of the mental degrees of freedom, $F^i \partial L = 0$. Then,

$$\dot{K} = G^a \partial_a H + J^i \partial_i L, \tag{27}$$

and the conservation of K leads to a relation between the coupling terms G^a and J^i: $G^a \partial_a H + J^i \partial_i L = 0$. If this condition holds, the Hamiltonian is part of a more general conserved quantity K. Since K is assumed to follow from the time-translation symmetry of the dynamics, its values should be identified with energy. Thus, L is a new form of energy associated to mental processes, and the notion of energy has to be extended in order to account for mind-body interactions.