The negative way to sentience

Ovidiu Cristinel Stoica *

March 23, 2020

Abstract

While the materialist paradigm is credited for the incredible success of science in describing the world, to some scientists and philosophers there seems to be something about subjective experience that is left out, in an apparently irreconcilable way. I show that indeed a scientific description of reality faces a serious limitation, which explains this position. On the other hand, to remain in the realm of science, I explore the problem of sentient experience in an indirect way, through its possible physical correlates. This can only be done in a negative way, which consists in the falsification of various hypotheses and the derivation of no-go results. The general approach I use here is based on simple mathematical proofs about dynamical systems, which I then particularize to several types of physical theories and interpretations of quantum mechanics. Despite choosing this scientifically-prudent approach, it turns out that various possibilities to consider sentience as fundamental make empirical predictions, ranging from some that can only be verified on a subjective basis to some about the physical correlates of sentience, which are independently falsifiable by objective means.

Contents

*Department of Theoretical Physics, National Institute of Physics and Nuclear Engineering – Horia Hulubei, Bucharest, Romania. Email: cristi.stoica@theory.nipne.ro, holotronix@gmail.com

1 With Galileo until the end (and beyond)

A common argument states that science can deal only with *quantities*, and consciousness requires something in addition – *qualities*. According to Chalmers, science can solve *easy problems of consciousness* (to explain what the brain *does* physically), but there is a *hard problem* (to explain *sentient experience*) (Chalmers, 1995). Scientists and philosophers of mind are divided, perhaps most of them thinking that there is no hard problem, or claiming to have an explanation how consciousness simply arises due to complexity, while others think that there is something irreducible, fundamental, about consciousness, and that the reductionist explanations can at best solve the easy problems. Philip Goff tracks this focus of science on quantities back to Galileo, and calls it *Galileo's error* (Goff, 2019). But was it an error, if science had no other way but to focus on the measurable? And if it was an error, why nobody fixed it since Galileo?

Here, I go in the opposite direction: rather than trying to fix "Galileo's error", I'll "persist in it", in order to push science to its own limits of explanatory power, to see if and where it breaks down. There are precedents – for example *Gödel incompleteness* (Gödel, 1931), Turing's noncomputability result (Turing, 1937), the *Bayes blind spot*, a severe limitation of Bayesian learning (Rédei and Gyenis, 2019), and Bell's theorem (Bell, 1964). So maybe we can find out that the materialist position cracks too. If so, and if consciousness has an irreducible, ineffable part, could it be possible to get some hints about it by studying the crack itself? Can the assumption that consciousness is fundamental even make some empirical predictions? I'll argue that the answer is yes.

I think the reason why many scientists claim that science is or will be able to explain all aspects of consciousness, is precisely because they are not going to the end with this reduc-

tionism to quantitative and measurable. Our human side contaminates this cold description of reality with anthropic intuitions and assumptions. This projection, due in part to our mirror neurons, gives the illusion that sentient experience can be reduced to the quantitative. To see this, rather than abandoning the quantitative, we will have to take it to the extreme, and clean science from the anthropic intuitions.

Despite focusing on the quantitative, science, empowered by mathematics, can go deeper in this direction, because

1. mathematics is the tool to identify the anthropic assumptions and to purify science, and
2. exact disciplines are often able to predict the limits of their own methods.

I will not discuss the "easy" problems and how they may be solved. They are not easy at all, and this subject is heavily researched in the fields of neuroscience and artificial intelligence. The "hard" problem on the other hand, was approached in various proposals in a very simple manner: bringing metaphysical arguments that there is something irreducible about consciousness, and elaborating intuition pumps to change the perspective on this (Dennett, 2013). But despite this, we seem to have no hint about what that irreducible part is and how it works with the physical world, which is the reason we call it "the hard problem". Nevertheless, I'll use a name for that irreducible part:

Nondefinition 1. *In the following, I will call* sentience *the ingredient that makes experience possible. Whatever this ingredient may be, I'll not try to define it.*

I leave undefined exactly what sentience is from physical, mathematical, or philosophical point of view, because all I know is that it makes experience possible. I'll take it as one, even if there are many sentient beings, because I am interested in its principles and laws, and how it relates to the observable world. It is similar to using the word "matter" even if we refer to the material bodies of many distinct objects and living beings. Whether sentience is an abstraction of consciousness or a concrete thing remains open for the moment. The fact that I leave it undefined should not be a problem, since I will focus on definable things only.

We will see that *science is unable to distinguish sentient beings, with their rich subjective experiences and behaviors, from timeless two-dimensional binary patterns.* This claim seems like a blasphemy which shouldn't remain unpunished. Before reacting like this, I invite the reader to go thoroughly through the proof (§2), and excuse the presence of mathematics, which I hope is present in tolerable doses. Even if the reader will still disagree with the conclusion, I hope that the least gains will be

1. a unified overview of science, in particular physics, in a nutshell,
2. a unified framework to analyze problems of philosophy and metaphysics,
3. and an insight into why some think that there is a *hard problem of consciousness* (Chalmers, 1995), while others find the idea ridiculous (Dennett, 2016; Churchland, 1981).

As a byproduct, this purification of science allows a clearer understanding of the underlying assumptions of the proposed solutions for the hard problem, and a clearer judgment of their tenability and empirical testability.

2 Science, purified

2.1 The physical world

It is an exaggeration that science deals only with quantities, it also deals with the qualitative, although not in the sense of the qualities of sentient experience. This is true also about the exact sciences, and about mathematics.

Principle 1. *Empirical data is about relations between objects, systems, and processes in the world, but not about their intrinsic nature.*

Evidence. Quantitative data is collected by counting and measurements. A measurement consists of comparing a property of an object or system with a similar property of an object or system called *etalon* (or *standard*). So the resulting quantitative data can be represented as numbers, but since they are ratios, they also represent relations.

Qualitative data is obtained by

1. Analyzing the structure of objects, how larger objects are composed of smaller objects.

2. Classifying objects and phenomena based on similarities and differences.

All these types of data represent relations, and say nothing about the intrinsic nature of the relata. □

Principle 2. *Laws and principles represent relations between empirical data.*

Evidence. A laws or a principle is a rule describing a regularity or a pattern in the raw empirical data. They can represent causal relations, structural patterns, statistical distributions or correlations *etc.* All these are relations between empirical data. □

Proposition 1. *A theoretical model is a description of relations. All it says about the relata are the relations in which they participate.*

Proof. Follows from Principles 1 and 2. □

The philosophical position that all we can know are structures, given as relations between abstract entities, is called *structural realism* (Ladyman, 2020).

Fortunately, there is a discipline which studies precisely the relations, and also helps identifying and solving the inconsistencies – *mathematics*. This is a central cause of the power of science. Mathematical structures are defined in terms of *universal algebra* (Grätzer, 2008). Accordingly

Definition 1. *A* mathematical structure *can be described as*

1. *A collection of* sets. *The nature of the elements of these sets is irrelevant.*

2. *A collections of* relations. *A relation is a subset of a Cartesian products of sets from our collection.*

One may wonder how could the nature of the elements of the sets from the mathematical structure be irrelevant. The reason is that all we need to know about the mathematical objects in order to make logically consistent mathematical proofs is contained in the collections of relations in which they participate. Definition 1 may seem too simple to cover the diversity and complexity of the mathematical structures, but in fact this is all that is needed, as proven by the fact that all mathematical formulations of various physical theories use mathematical structures that can be defined like this.

Example 1. *An order relation $\leq$ on a set X is the subset of $X \times X$ consisting of the pairs of the form $(x, y) \in X \times X$ satisfying $x \leq x$. A function $f : X \to Y$ can be specified by its graphic, which is a subset of $X \times Y$. A binary operation $+$ on X can as well be specified by its graphic, which is the set of all $(x, y, x+y) \in X \times X \times X$. Vector spaces, topology, differentiable manifolds, they all can ultimately be formulated in terms of sets and relations.*

Definition 1 is so general, that it can be used to build mathematical models of axiomatic systems, as it is done in *model theory* (Chang and Keisler, 1990; Hodges, 1997). If we collect all propositions that we found to be true about the world, they should not contradict one another. Together they form a *theory*, which can be modeled by a mathematical structure.

Therefore, the focus of science on relations gave it a great advantage, by making it perfectly suited for mathematical formulations. This could explain its unreasonable effectiveness (Wigner, 1960; Tegmark, 2014; Stoica, 2015b), and justify Galileo's remark, *"The book [of Nature] is written in mathematical language"*.

In the following, I will assume this as a principle.

Principle 3. *Any consistent theory about reality admits a mathematical model.*

Argument. The relation between theories and mathematical models is expressed by Henkin's *model existence theorem* (Henkin, 1949), stating that

If a first-order theory is logically consistent, then it admits a mathematical model.

Gödel's *completeness theorem* (Gödel, 1930) states that any first-order theory is complete, in the sense that any logically valid formula can be obtained by a finite deduction from the axioms. Henkin found a simplified, constructive proof of the completeness theorem, in terms of the model existence theorem (Henkin, 1949). Then he obtained a generalization of the model existence theorem to second-order theories (Henkin, 1950, 1996). Various results beyond first-order theories exist, but there are also counterexamples. Despite the existence of counterexamples, no consistent theory in natural sciences is known to have no mathematical model. This is why I will take this as a principle. □

In other words, I will assume that the existence of a mathematical model is a prerequisite for physical existence.

I will not be interested whether or not a theory is Gödel complete, as long as it admits a mathematical model, since for the purpose of this essay I am interested in the model only, and not in its possible axiomatizations. The model may satisfy propositions that can't be proven by finite length proofs from a particular set of axioms, but axiomatization is just a way to organize the truths about the mathematical structure by reducing them to a finite set of axioms, which is not the purpose of this essay.

2.2 Dynamical systems

It requires a serious abstraction effort to think of theories as mathematical structures in general. In fact, since time plays an important role in natural sciences, we can particularize to those mathematical structures that are dynamical systems, simplifying our discussion. To see this, let's consider a very simple example.

Example 2 (Free classical particle)**.** *Let's consider a classical particle in the one-dimensional space* $\mathbb{R}$ *(the set of real numbers). Suppose it is free – no force acts on it. Then, according to* Newton's first law of motion*, the particle is either at rest, or it moves with constant velocity. The equation of motion of the particle is*

$$x(t) = x_0 + vt, \tag{1}$$

where the particle's position at time t *is* $x(t)$*,* $x_0 = x(t_0)$ *is it's initial position at time* $t_0 = 0$*, and* v *is its constant velocity. If* $v = 0$ *the particle stays at rest at the same position* $x(t) = x_0$*.*

The configuration space *is the space of all possible positions, in this case* $\mathbb{R}$*. But just knowing the initial position of the particle* x_0 *is not enough to find out* $x(t)$*, unless we also know the*

velocity v. *The pair* (x, v) *is enough to determine all future (and past) motion of the free particle. Let's call* $s = (x, v) \in \mathbb{R}^2$ *the* state *of the particle. All possible states can be collected in the* state space $\mathbb{R}^2$. *The equation of motion* (1) *tells us that the state of the particle,* $s(t) = (x(t), v)$ *is always on the line from the state space* $\mathbb{R}^2$ *which is determined by the condition that* v *is constant and passes through* (x_0, v) *(Fig. 1). A common equivalent representation is the* phase space *representation, in which instead of velocities* v *one uses momenta,* mv, *where* $m \neq 0$ *is the particle's mass. In the following I will use the term* state space *in all cases.*

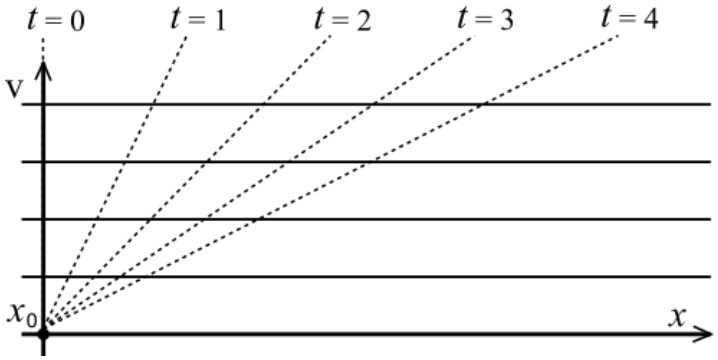

Figure 1: The dynamical system from Example 2. The points represent states (x, v). The trajectory in the state space for a system starting in (x_0, v) is a horizontal line for $v \neq 0$, and a point at the origin $(x_0, 0)$ for $v = 0$. Time is a parameter along each of these trajectories.

The main idea from Example 2 is that the entire past and future history of the system is determined by the state at a given time and the law of motion. Surprising as it may seem, all known deterministic theories in physics are like this (Fig. 2 A). Whether we talk about more particles, forces, fields, fluids, quantum particles, quantum fields, there is always a state space, and the history of the particle is a curve in that space. Time is just a parameter on the curve. If the theory is not deterministic, the difference is that the path of the system can branch and is not uniquely determined by the initial state (Fig. 2 B). If the evolution consists of a deterministic law combined with probabilistic jumps, like standard quantum mechanics, the trajectory can be discontinuous in the state space.

Definition 2. *A* deterministic dynamical system *(*eg. *Meiss (2007)) a* $\mathcal{A} = (S, T, \tau)$ *consists of*

1. *a* state space *(or* phase space*)* S,
2. *a one-dimensional group or monoid* T, *which can be the real numbers* $\mathbb{R}$ *or the integers* $\mathbb{Z}$, *or the non-negative reals* $\mathbb{R}^+$ *or integers* $\mathbb{N}$,
3. *an* evolution law $\tau : T \times S \to S$, *which satisfies* $\tau(t_2, \tau(t_1, s)) = \tau(t_1 + t_2, s)$, $\tau(0, s) = s$, *for all* $t_1, t_2 \in T$ *and* $s \in S$.

The possible states of the system are represented as points in the state space S, but they are in fact mathematical structures. Space is included in these structures, as well as the distribution of particles or fields in space. In the case of quantum mechanics, the wavefunction is defined not on space, but on the Cartesian product of n copies of space with itself, where n is the number of particles. Usually it is possible to endow the state space S itself with a structure which contains all the information about the structure of its states.

Definition 3. *The state space* S *contains all possible states, but the actual state of a system is a single element* $s_0 \in S$. *The* history *of the system starting in the state* s_0 *is a function* $\gamma_{s_0} : T \to S$, $\gamma_{s_0}(t) := \tau(t, s_0)$. *In this case,* s_0 *is called* initial condition, *and the unique path* $\gamma_{s_0}(T) = \tau(T, s_0)$ *is a* solution *of the evolution law with the initial condition* s_0.

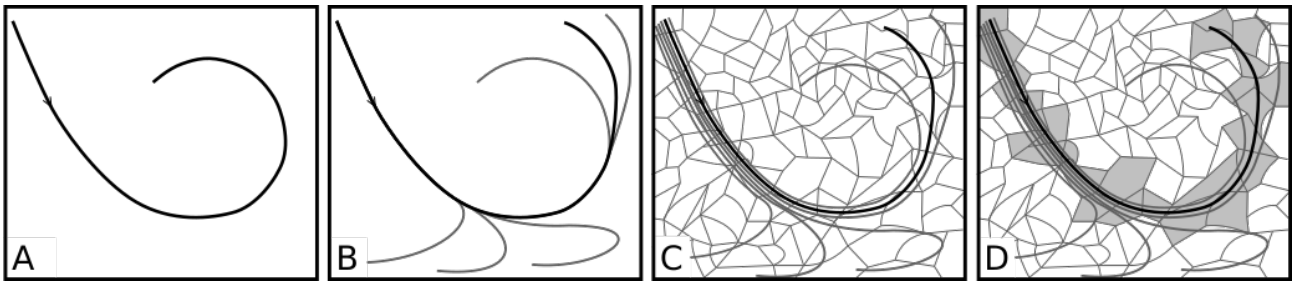

Figure 2: Dynamical systems. **A.** Deterministic. **B.** Nondeterministic. **C.** Coarse graining of a deterministic dynamical system. **D.** The coarse graining of a deterministic system can be nondeterministic.

Example 3. *All theories in classical physics are deterministic dynamical systems. To describe as dynamical systems special and general relativity, which don't have a preferred time, we need to choose one, and use a time formulation like (Arnowitt et al., 2008). Quantum mechanics and quantum field theory are deterministic dynamical systems as long as no collapse of the wavefunction is involved.*

Example 4. *If the set S in Def. 2 is discrete, T is also discrete, and the system is a* discrete state machine*, proposed by Alan Turing – what is called today* Turing machine *(Turing, 1950). Other examples of discrete dynamical systems are* finite state machines *and* cellular automata*.*

There are more possible ways to modify Definition 2 to include nondeterministic theories. Here I will give a simple one.

Definition 4. *A* (nondeterministic) dynamical system $\mathcal{A} = (S, T, \tau)$ *consists of*

1. *a* state space S*,*
2. *a one-dimensional group or monoid T as in Definition 2,*
3. *an* evolution law $\tau : T \times S \to \mathscr{P}(S)$*, where $\mathscr{P}(S)$ is the set of all subsets of S, and τ satisfies the conditions $\tau(t_2, \tau(t_1, s)) = \tau(t_1 + t_2, s)$, $\tau(0, s) = \{s\}$, for all $t_1, t_2 \in T$ and $s \in S$.*

At need, Def. 4 can be completed by assigning probability distributions on the sets $\tau(t, s)$, and imposing compatibility conditions.

Definition 5. *For a nondeterministic system, a history starting in the state s_0 is also a function $\gamma_{s_0} : T \to S$, so that $\gamma_{s_0}(t) \in \tau(t, s_0)$, and $\gamma_{s_0}(t_1 + t_2) = \gamma_{\gamma_{s_0}(t_2)}(t_1)$ for all $t, t_1, t_2 \in T$. Unlike in the deterministic case, the history is not uniquely defined by s_0.*

Example 5. *If $\mathcal{A} = (S, T, \tau)$ is a nondeterministic dynamical system such that $\tau(t, s)$ has only one element for all $t \in T$ and $s \in S$, then it is a deterministic dynamical system (Def. 2). For this reason, from now on we will use the term* dynamical system *according to Def. 4.*

Example 6. *If the state space S from Def. 4 is discrete, we get* nondeterministic Turing machines*.*

Example 7. *In quantum mechanics, the state space is a Hilbert space $\mathcal{H}$, and the one-dimensional group is the subgroup of the unitary group of $\mathcal{H}$ generated by the Hamiltonian of the system. When a measurement takes place, in general, the evolution law is considered to no longer obey the Schrödinger equation for that moment, and includes discontinuous jumps, resulting in multiple possible states (von Neumann, 1955).*

The richness of dynamical system allows us to admit the following:

Principle 4. *Any consistent physical theory can be modeled by a dynamical system.*

2.3 Subsystems

The notion of dynamical system hides the structure of the states, treating them as points in the state space. But in many cases this structure is relevant. It is possible to give a general treatment like the one using dynamical systems presented here, which also takes into consideration the detailed structure of the states (Stoica, 2008a), but, for the purpose of this discussion, I tried to reduce this at minimum, for simplicity. Here I will introduce only a minimal complication: since each state is a mathematical structure of a certain type, they are *objects* in the category $\mathcal{C}$ of that type of mathematical structures.

Since the states from S are mathematical structure of a specific type, in general the notion of substructure of that type makes sense. In general, a substructure of a mathematical structure from a given category is defined as a subset such that the restrictions of the relations and operations still satisfy the definition of the objects of the same category.

Definition 6. *A* category $\mathcal{C}$ *consists of a class of* objects $\mathrm{ob}(\mathcal{C})$*, (which for our interest are mathematical structures of a particular type), and a collection of* morphisms $\mathrm{hom}(\mathcal{C})$ *between the objects (Mac Lane, 1998). The operation of* composition $\circ$ *of morphisms is associative and has an identity* 1_A *for each object* $A \in \mathrm{ob}(\mathcal{C})$*. A* monomorphism *is a morphism* $\in \mathrm{hom}(\mathcal{C})$ *such that* $f \circ g_1 = f \circ g_2 \Rightarrow g_1 = g_2$ *for all morphisms* g_1, g_2*. If* $f, g, h \in \mathrm{hom}(\mathcal{C})$ *such that* $f = g \circ h$*, we write* $f \leq g$*. We write* $f \equiv g$ *if* $f \leq g$ *and* $g \leq f$*. The relation* $\equiv$ *is equivalence relation on the monomorphisms with codomain* $A \in \mathrm{ob}(\mathcal{C})$*, whose classes are called* subobjects *of* A*. The* dual *or* opposite *category* $\mathcal{C}^{\mathrm{op}}$ *is obtained from* $\mathcal{C}$ *by reversing the morphisms and their composition. An* epimorphism *in* $\mathcal{C}$ *is defined as the dual of a monomorphism in* $\mathcal{C}^{\mathrm{op}}$*. A* quotient *object in* $\mathcal{C}$ *is defined as a subobject in* $\mathcal{C}^{\mathrm{op}}$*.*

Definition 7. *A system* $\mathcal{S} = (S', T', \tau')$ *is called* subsystem *of a system* $\mathcal{A} = (S, T, \tau)$ *if the states from* S' *are subobjects or representatives of subobjects of the states of* S *in the same category* $\mathcal{C}$*.*

Definition 8. *For some categories, a subobject* s' *of an object* s *has a* complement *in* s*, which is also a subobject. If this is the case for the category in Def. 7, then the subsystem* $\mathcal{S}$ *of* $\mathcal{A}$ *determines a complement subsystem, called the* environment *of the system* $\mathcal{S}$*. In categories with complemented subobjects, objects can be seen as composed of simpler objects. In* monoidal categories*, which incorporate the* tensor product *of objects, there are objects that can't be decomposed like this.*

Example 8. *A subsystem of a system of* n *classical particles is obtained by considering a subset of* k *particles out of the* n *particles, where* $k \leq n$*. Its environment is the complement subsystem, consisting of the remaining* $n - k$ *particles. An example of dynamical systems whose states are not decomposable as in Def. 8 is given by the entangled systems in quantum mechanics.*

Definition 9. *A subsystem is called* closed *or* isolated *if it is independent of the environment, otherwise it is called* open*.*

3 Reductionism

3.1 Definitions and examples

Definition 10. *A dynamical system* $\mathcal{B} = (S_{\mathcal{B}}, T_{\mathcal{B}}, \tau_{\mathcal{B}})$ *can be* reduced *to a dynamical system* $\mathcal{A} = (S_{\mathcal{A}}, T_{\mathcal{A}}, \tau_{\mathcal{A}})$ *if there are two maps,*

1. *a morphism of monoids* $f_T : T_{\mathcal{A}} \to T_{\mathcal{B}}$*, and*

2. *a* reduction map $f_S : S_{\mathcal{A}} \to S_{\mathcal{B}}$,

satisfying

$$\tau_{\mathcal{B}}(t_{\mathcal{B}}, s_{\mathcal{B}}) = \bigcup_{(t_{\mathcal{A}}, s_{\mathcal{A}}) \in f_T^{-1}(t_{\mathcal{B}}) \times f_S^{-1}(s_{\mathcal{B}})} f_S\left(\tau_{\mathcal{A}}(t_{\mathcal{A}}, s_{\mathcal{A}})\right). \tag{2}$$

The state space $S_{\mathcal{B}}$ is called coarse graining *of the state space $S_{\mathcal{A}}$, because its states correspond to equivalence classes of states from $S_{\mathcal{A}}$, defined by $[s_{\mathcal{A}}]_{f_S} := f_S^{-1}\left(f^S(s_{\mathcal{A}})\right)$.*

A state $s_{\mathcal{B}} = [s_{\mathcal{A}}]_{f_S}$ from $S_{\mathcal{B}}$ is called macro-state*, and a particular representative $s_{\mathcal{A}}$ from $S_{\mathcal{A}}$ is called* micro-state *of $s_{\mathcal{B}}$ (Fig. 2 C).*

The evolution law $\tau_{\mathcal{B}}$ is called independent of the micro-states *if*

$$\tau_{\mathcal{B}}(t_{\mathcal{B}}, s_{\mathcal{B}}) = f_S\left(\tau_{\mathcal{A}}(t_{\mathcal{A}}, s_{\mathcal{A}})\right) \tag{3}$$

for all $(t_{\mathcal{A}}, s_{\mathcal{A}}) \in f_T^{-1}(t_{\mathcal{B}}) \times f_S^{-1}(s_{\mathcal{B}})$.

Since each state is also a mathematical structure, the reduction map in Def. 10 is in general either a morphism of universal algebras which reduces some of the relations to trivial ones, or a forgetful functor.

Definition 11. *With the notations from Def. 10, the system $\mathcal{B}$ is said to* reduce *to the system $\mathcal{A}$, or to be an* abstraction *of $\mathcal{A}$. The properties of a state $s_{\mathcal{B}} \in S_{\mathcal{B}}$ which are not among the properties of $s_{\mathcal{A}} \in f_S^{-1}(s_{\mathcal{B}})$ are called* abstract *properties* [1].

A property of the system $\mathcal{B}$ is called weakly reducible *if is Gödel undecidable (Gödel, 1931) or Turing uncomputable (Turing, 1937) from the properties of the system $\mathcal{A}$, otherwise it is called* strongly reducible.

Example 9 (Thermodynamics)**.** *An example of success of reductionism is the reduction of classical thermodynamics to classical statistical mechanics. The macro states in thermodynamics correspond to coarse graining regions of the microscopic state space of classical mechanics. The logarithm of the volume in the state space of the coarse graining regions gives the Boltzmann entropy. Properties like temperature and heat capacity appear in thermodynamics as aggregate properties of the microscopic states.*

Example 10 (Quantum to classical)**.** *By contrast, the macro states can't be reduced to quantum states as in Def. 10. Call the (micro) quantum states Q-states. For each macro state, there is at least a Q-state to which it reduces, call such Q-states C-states. But there are Q-states that don't correspond to observable macro states* – eg. *macro quantum superpositions similar to Schrödinger's cat – call them $\not C$-states. The existence of $\not C$-states is the reason why the fact that the world appears classical at the macro level is a problem in quantum mechanics, and is also part of the* measurement problem*. One possible solution could be to remove the $\not C$-states from the quantum state space, to make the reductionist recipe from Def. 10 work. Unfortunately C-states may evolve into $\not C$-states. Fortunately, C-states that evolve only in C-states by the Schrödinger equation (without collapse) may exist, so it is indeed possible to limit the quantum state space to only such states (Schulman, 1997; 't Hooft, 2016; Stoica, 2016a, 2017). Other solutions are (Everett, 1973), where* decoherence *is expected to solve the problem by branching the evolution (the* many-worlds interpretation *or* MWI*), and in (Bohm, 1952), which includes, along with the branching wavefunction representing the quantum state, point-particles which select one of the branches. Another way is to change the evolution law, so that as soon as a C-state evolves into a $\not C$-state, it collapses back to a C-state, as in (Ghirardi et al., 1986).*

But there is another problem with the macro picture, since the experimental violation (see Leggett (2002); Emary et al. (2013) and references therein) of the Leggett-Garg inequality

[1] I am trying to avoid the use of the term *emergence*, because it is used with opposite meanings by various people.

shows that in a quantum world the macro level having a definite macro state can be in conflict with the ability to determine that state by using noninvasive observations (Leggett and Garg, 1985; Leggett, 2008).

Remark 1. *There is a widespread consensus among scientists that the reductionism program is successful, and it's only a matter of time and complexity until all sciences will be reduced to fundamental physics. However, Example 10 shows that even for physics this can be problematic.*

Example 11 (Subsystem)**.** *The Example 8 of a subsystem of classical particles suggests that a subsystem is at the same time the dynamical system obtained by ignoring the other $n-k$ particles. In other words, we stop distinguishing between the states in which the k particles have the same positions and velocities, disregarding the positions and velocities of the other $n-k$ particles. Hence, a subsystem satisfies Def. 10.*

Proposition 2. *For any nondeterministic dynamical system $\mathcal{A}$, a deterministic one $\mathcal{A}'$ can be constructed in theory, so that $\mathcal{A}$ is reducible to $\mathcal{A}'$.*

Proof. Let $\mathcal{A} = (S, T, \tau)$ be a nondeterministic dynamical system. Define S' as the set of all the pairs (s_0, γ_{s_0}), where $s_0 \in S$, and γ_{s_0} is a history starting in the state s_0 *cf.* Def. 5. Define $\tau' : T \times S' \to S'$ by $\tau'(t, (s_0, \gamma_{s_0})) = (s, \gamma'_s)$, where $s \in \tau(t, s_0) \cap \gamma_{s_0}$, and γ'_s is the only history starting in the state s which is included in γ_{s_0}. Then, the dynamical system $\mathcal{A}' = (S', T, \tau')$ is deterministic, and $\mathcal{A}$ is reducible to $\mathcal{A}'$ *cf.* Def. 10 (Fig. 2 C and D). □

Corollary 3. *Any nondeterministic Turing machine is reducible to a deterministic one.*

Proof. With the notation in the proof of Prop. 2, since S, T, and the set of possible histories are discrete, it follows that the deterministic system $\mathcal{A}'$ is also discrete. □

Corollary 4. *Quantum mechanics can be reduced to a deterministic theory.*

Proof. Follows from Prop. 2. A known example is the *pilot-wave theory* (Bohm, 1952), where the collapsing wavefunction (von Neumann, 1955) corresponds to the conditional wavefunction, and the coarse graining consists of ignoring the "hidden variables" (positions in space). □

Definition 12. *Let $\mathcal{A} = (S, T, \tau)$ be a dynamical system, and $\mathscr{S} = (S_\alpha)_{\alpha \in \mathscr{A}}$ a partition of S,* i.e. *$S = \bigcup_{\alpha \in \mathscr{A}} S_\alpha$ and $S_\alpha \cap S_\beta = \varnothing$ for all $\alpha, \beta \in \mathscr{A}$, $\alpha \neq \beta$. Then, the evolution rule on $\mathcal{A}$ induces an evolution rule τ' on $\mathscr{S}$, and a partition T' of T, resulting in a dynamical system $\mathcal{A}' = (\mathscr{S}, T', \tau')$. Thus, by Def. 10, $\mathscr{S}$ is a coarse graining of S, and $\mathcal{A}'$ is reducible to $\mathcal{A}$. We say that $\mathcal{A}'$* approximates *$\mathcal{A}$. In particular, if the state space S is continuous and the index set $\mathscr{A}$ is discrete, T' is also discrete, and $\mathcal{A}'$ is called* discretization *of $\mathcal{A}$.*

Example 12 (Simulation)**.** *Suppose that a dynamical system $\mathcal{B}$ is a discretization of a dynamical system $\mathcal{A}$. Since $\mathcal{B}$ is discrete, it is a (possibly nondeterministic) Turing machine,* cf. *Examples 4 and 6. Therefore, $\mathcal{B}$ is a* simulation *of $\mathcal{A}$.*

Remark 2. *In general, an object is isomorphic with subobjects of more different objects. Therefore, a dynamical subsystem (Def. 7) is at the same time an approximation of the system $\mathcal{A}$ (Def. 12), and therefore is also reducible to $\mathcal{A}$ (Def. 10).*

Definition 13. *We call* implementation *of a dynamical system $\mathcal{A}$ any realization of $\mathcal{A}$ as an abstraction (*cf. *Def. 11) of a physical system.*

Example 13. *Consider a finite state machine $\mathcal{F}$ which is can be implemented on a physical subsystem $\mathcal{K}$. For example, $\mathcal{K}$ can be a classical computer – the* hardware *of $\mathcal{F}$, and $\mathcal{F}$ a* software*. Because a computer has input devices (and also because you can break it), it's an open system. While classical computers we use have components that employ quantum mechanical effects, they can be made to use only classical mechanics.*

3.2 Time

For a system $\mathcal{A} = (S, T, \tau)$, space is implicit in the structure of the states from S, but time is explicit in the parameter space T. This is true also for relativistic spacetime, in dynamical formulations like (Arnowitt et al., 2008). But the system $\mathcal{A} = (S, T, \tau)$ itself is just the state space S, whose states are connected to one another by the evolution law τ (a binary relation on S), time being just a parameter on the resulting curves in S representing the histories of the system. So in fact, despite including time as a parameter and being called "dynamical", dynamical systems are static mathematical structures. Nothing changes in them.

Definition 14. *The view that the present time is the only real is called* presentism. *The view that all moments of time are equally real and the past, present, and future are relative notions is called* eternalism.

Maybe a presentist reader had no problem with this so far, because he or she could imagine that the finger of a being outside the dynamical system itself is pointing the present state of the system among the other points of the state space S. But there is nothing in the system that needs such a finger. There is nothing in the data to highlight a present time. All we have are temporal relations. *Present is a notion relative to the state of the system*, so for each state s in S, the present time is when the system is in the state s. If the state s includes human beings, it includes as well their memories, plans for future, and neural correlates of thoughts like "it's five o'clock" or "I feel happy about this memory". Therefore, the histories of dynamical systems are as *eternalist* as the block world of the theory of relativity. This is true even for nondeterministic systems, for each possible history of the system. Any dynamical system, and each history of it, are purely statical structures.

Despite the absence of such a "finger" or "pick-up needle" that would give the present a special status, one may think that there is some sort of mechanism necessary to "run" the dynamical system in a temporal manner. For example, if an algorithm $\mathcal{F}$ is a Turing machine, that you would need a computer $\mathcal{K}$, another Turing machine, more "real", to run it. But this would just leave us in the same place, since that hardware, no matter how "hard" $\mathcal{K}$ is, it's still a dynamical system, and even if it has a processor clock, there is nothing physical attached to the present tick of the clock to make it more "real" than any other tick. The computer $\mathcal{K}$ just implements, in the sense of Def. 13, the software $\mathcal{F}$, and the software is an abstraction of $\mathcal{K}$ (Def. 11), so it's reducible (Def. 10) to $\mathcal{K}$, but nothing to make time more than a parameter happens in either $\mathcal{K}$ or $\mathcal{F}$.

Proposition 5. *There is no empirical or theoretical scientific evidence for presentism.*

Proof. This follows from Principle 4. More precisely, we can only access time relations, but not time itself, both empirically (Principle 1) and theoretically (Principles 2 and 4). □

This may seem unconvincing, so I'll give an explicit proof.

Theorem 6. *Trying to complete a dynamical system with a presentist notion of time results in yet another dynamical system, which is still static, but more complicated.*

Proof. Let's try to make the present time "physical". Consider a dynamical system $\mathcal{A} = (S, T, \tau)$, where T is the one-parameter group $\mathbb{R}$ or $\mathbb{Z}$. We can try to label each state with an element of a set $\mathscr{T}$, which can be $\mathscr{T} = \{-, 0, +\}$, where each of the elements means that the state is respectively in the past, present, and future (or possible future), or $\mathscr{T} = \{0, 1\}$, meaning "real" (present, actual), respectively "unreal" (past or future or possible future), or we can even take $\mathscr{T} = T$ to label the time interval separating the past or future states from the present, the proof is the same. Then, we have to modify the state space S so that each state can be in addition

in one of the possible states from $\mathscr{T}$, so we replace a system S by $S' = S \times \mathscr{T}$. Hence, instead of states $s \in S$, we will have pairs of the form (s, θ) with $s \in S$ and $\theta \in \mathscr{T}$. Now consider in $\mathcal{A}$ a history γ_{s_0} where $s_0 \in S$, *cf.* Def. 5. At the time $t = 0$, s_0 must be labeled as present, so we replace it with $(s_0, 0) \in S'$. Since the states $s \in \gamma_{s_0}$ are of the form $s = \tau(t, s_0)$, we replace them by $(s, +)$ if $t < 0$, and with $(s, -)$ if $t > 0$. Let $\gamma'_{s_0} = \{(\tau(t, s_0), \text{sign}(t)) \in S' | t \in T\}$, where sign is the sign function. Now we see that, when time changes the original state from s_0 to $s = \tau(t, s_0)$, the curve γ'_{s_0} is replaced by a curve $\gamma'_s = \{(\tau(t, s'), \text{sign}(t)) \in S' | t \in T\}$. Since T is a group, $\gamma_{s_0}(T) = \gamma_s(T)$, so the two curves γ'_{s_0} and γ'_s differ only in the θ components. What we achieved so far is just that instead of states $s \in S$, we now have to use the entire histories γ'_s as states, since not only $s' = (s, \theta) \in \mathscr{S}'$ changes, but also the past and future of each s'. So we can't take as state space S', but the collection of histories in S'. If we do this, we get a dynamical system $\mathcal{A}' = \{\widetilde{S}, T, \widetilde{\tau}\}$, where $\widetilde{S}$ is the collection of histories in S', and the evolution law $\widetilde{\tau}$ is defined by $\widetilde{\tau}_t(\gamma'_s) = \gamma'_{\tau(t,s)}$. Not only we obtained again a dynamical system, but now the states have to include, in addition to the present state, the past and future states as well. This is not an improvement compared to the block world view, it is in fact worse, since now we have a different block world at each instant, and all these block worlds are combined in a hyper-presentist way in a higher order block world. □

So, if we want to take the presentist position, the model gets complicated, but it still remains a dynamical system with no preferred present. If we still insist, the only thing we can do is to add presentism as a metaphysical assumption. This is not justified by scientific evidence, but maybe it is justified by our *experience of time* as flowing, which seems opposite to the idea of block world. But I think the keyword here is not "time", but "experience".

3.3 Materialism

Materialism is sometimes presented as the direct cause of the success of science. But, if Proposition 1 is true and we only know the relations, how can we know that the relata is matter, or even if it exists in a stronger sense than its relations?

To see this more clearly, we can do the following exercise. Imagine that we would ask a materialist from 150 years ago what would think about lengths and durations appearing to be different in different reference frames, and time being the fourth dimension. Probably the reply would be "have you lost your marbles like Zöllner (Zöllner, 1880)?"

But soon special relativity was discovered, followed by the even more radical general relativity. This made Bertrand Russell to say that materialism is no longer supported (Russell, 1921)

> *Whoever reads, for example, Professor Eddington's "Space, Time and Gravitation" (Cambridge University Press, 1920), will see that an old-fashioned materialism can receive no support from modern physics. I think that what has permanent value in the outlook of the behaviourists is the feeling that physics is the most fundamental science at present in existence. But this position cannot be called materialistic, if, as seems to be the case, physics does not assume the existence of matter.*

What would a materialist from 100 years ago have to say about reality seeming to depend on the experiment you perform? Or, assuming the speed limit from relativity, what would she or he say about the idea of correlations at a distance, as in the experiments with entangled particles? Probably that these are crazy ideas, and only a deluded occultist may believe them. Yet, quantum mechanics appears to be just like this. And there's no way to fix it to fit the taste of mainstream materialism.

Is physics about material stuff which is in a definite state, independent of the observer, and which doesn't break the speed of light? Bell's theorem (Bell, 1964, 2004) and the experimental confirmations of violations of Bell's inequality force us to make an impossible choice – between

accepting spooky action at a distance, and accepting that particles don't have a definite state independent of the experiments that will be performed in the future. So we have to choose between what Einstein called "spooky action at a distance" (nonlocality in space) and something that looks like predestination, conspiracy, or retrocausality (nonlocality in time). Some try to opt out of this, by denying reality itself. Others try to do it by accepting that all outcomes are realized in different worlds. There is no consensus among physicists, because any choice means to accept some crazy idea. All of these options fly in the face of mainstream materialism.

So the following proposition, while true, is an understatement:

Proposition 7. *There is no empirical or theoretical scientific evidence for matter.*

Proof. Science can only tell us about relations, not about relata (Proposition 1). □

This doesn't mean that there is no matter, just that science can't tell us whether the relations we can access empirically or theoretically are between matter or something else, or whether there's anything about the relata more than the relations they take part in.

Materialism can't be disproved not because it's true, but because whatever we discover, regardless how different it is from what materialists used to say matter is, we will call it matter. But science can only tell the relations, and nothing about the nature of the relata. The idea of matter comes perhaps from our sensory experiences with the world, which we generalized unjustifiably to the nature of the elementary building blocks of the world. But this idea of matter is a projection onto the microscopic, fundamental world, of our own sensory experiences of the macro world. Matter as we know it exists only in our mental and sensory representations.

Instead of stretching the meaning of the word "materialism" to cover any situation, a better name for the attitude accompanying modern physics is *physicalism*. But even this term, which means "everything is physical", is used sometimes as implying a sort of upgraded materialism, and makes us forget or ignore the fact that science can only tell the relations, and nothing about the relata.

3.4 Ontology

Since science, like all of our objective knowledge, deals with relations only, it appears that it can say nothing about the relata. Questions about "the nature of things" or "being", rather than merely about the relations, about "what is" rather than "how it behaves" or "what processes happen", seem to remain unanswerable, at least by objective methods. *Ontology* is supposed to deal with them, but ontology is metaphysical in nature. Whatever we can say about the nature of existence are metaphysical assumptions, or not even that, since they are projections of our sensory experiences.

Yet, we can at least hope to isolate the place where ontology impacts the physical-relational description, *i.e.* its *physical correlate*, by using the very methods of science. To do this, we need to clarify what to expect from ontology. The definitory requirement is that all that exists physically is grounded on the ontology, that the states and processes of the ontological level completely determine the states and processes of the physical world.

Definition 15. *Let $\mathcal{P} = (S_\mathcal{P}, T_\mathcal{P}, \tau_\mathcal{P})$ be the dynamical system of the physical world, and $\mathcal{O} = (S_\mathcal{O}, T_\mathcal{O}, \tau_\mathcal{O})$ another system, so that the states of both systems are are objects in the categories $\mathcal{C}(S_\mathcal{P})$, respectively $\mathcal{C}(S_\mathcal{O})$. Then, if $\mathcal{O}$ reduces to $\mathcal{P}$ (*cf. *Def. 10) via a* fully faithful forgetful functor *(Mac Lane, 1998) between the categories $\mathcal{C}(S_\mathcal{P})$ and $\mathcal{C}(S_\mathcal{O})$, we call the system $\mathcal{O}$* physical correlate of ontology candidate *(PCOC) or a* foundation *of the system $\mathcal{P}$. If $\mathcal{P}'$ is another system which is reducible to $\mathcal{P}$, and if $\mathcal{O}$ is a foundation of $\mathcal{P}$, we also say that $\mathcal{O}$ is a foundation of $\mathcal{P}'$.*

A foundation $\mathcal{O}$ of $\mathcal{P}$ is called irreducible*, or a* basis *of $\mathcal{P}$, if for any other foundation $\mathcal{O}'$ of both $\mathcal{P}$ and $\mathcal{O}$, $\mathcal{O}$ is a foundation of $\mathcal{O}'$.*

Remark 3. *The trivial foundation of the system $\mathcal{P}$ is, of course, $\mathcal{P}$ itself. But the Definition 15 is not empty, since the states of $\mathcal{P}$ may have relations or properties that can be derived from other relations or properties. In this case, which is quite general, the functor in the definition forgets some of the relations in a way which allows them to be "remembered" or reconstructed from its relations, so the foundation is a more parsimonious system which still captures all the relations or properties and the dynamics of $\mathcal{P}$.*

It may seem counterintuitive that Def. 15 assumes the reduction of $\mathcal{O}$ to $\mathcal{P}$, rather than the other way around. The reason is that, since the functor is fully faithful, $\mathcal{P}$ also reduces to $\mathcal{O}$. The key difference is that the functor is forgetful, and this makes $\mathcal{O}$ more fundamental than $\mathcal{P}$.

If $\mathcal{O}$ is irreducible, this is to say that it is minimal or simpler. The goal of reductionism, and in fact of the unification programs, is to get not merely any system $\mathcal{P}$ which faithfully corresponds to the world, but to also be irreducible. This is in fact a quest for ontology, in the limited sense in which it can be reached, i.e. *as PCOC. Also note that $\mathcal{P}$ may admit more than one irreducible foundations.*

Example 14. *Classical mechanics is a foundation of thermodynamics, since the latter reduces to the former* cf. *. Example 9). Since all classical observables are functions of the generalized coordinates q_j and generalized momenta p_j in the Hamiltonian formulation, a description in terms of q_j and p_j alone is a PCOC or a foundation of classical mechanics.*

An example of a theory which doesn't have a proper foundation yet is given by Example 10. However, quantum mechanics itself, assuming there is no collapse, has a PCOC, which is the wavefunction, since all observables can be obtained from it. Since the collapse seems to be necessary during measurements, this foundation seems to be insufficient, and the task of the interpretations of quantum mechanics is also to provide a better one.

We see that Def. 15 doesn't say anything about the relata, it only identifies physical correlates of ontology candidates.

It is sometimes said that physics tell us nothing about *being*, only about *doing*. I think it's more appropriate to say that physics tells us only about *being in relation to*, but not about *being in itself*.

4 Is sentience reducible?

4.1 A hard problem

Even if both matter and information are human constructs, the possibility that consciousness is fully reducible to physical processes seems to remain open. This relies on *functionalism* – the thesis that there is nothing more about consciousness than it's states and processes and the way it functions.

Theorem 8. *Any theoretical model of a human is equivalent to a Turing machine.*

Proof. Consider a dynamical system describing the world, including humans as subsystems. Since the human brain occupies a finite volume in space, quantum mechanics says that it can only have a discrete number of possible observable states. But even if this was not the case, in general we can assume that we can approximate it by a discrete dynamical system. While the brain is a sophisticated neural network, it's also a discrete state machine, *i.e.* a Turing machine. At the macro level it may seem continuous, but at micro level is discrete because as a quantum system it is bounded in space. The same is true about human's entire body. The rest of the world, the environment of the human, is not bounded in space, but we are interested in it's inputs to the human. Since the human senses have limited resolution, those inputs can as well be discretized. So, the human and the relevant environment can be discretized without

relevant loss, and from this point of view they can be modeled by a nondeterministic Turing machine. From Corollary 3, the Turing machine can be chosen to be deterministic, or reduced to a deterministic one. □

Theorem 9. *Any theoretical model describing a human is equivalent to a two-dimensional timeless tapestry generated by the* Rule 110 *(Fig. 3).*

Proof. Rule 110 is a cellular automaton. In 1985, Stephen Wolfram conjectured that it is *Turing complete*, *i.e.* capable to simulate any Turing machine. This was proven by Cook (Cook, 2004). By combining with Theorem 8, we get the proof. □

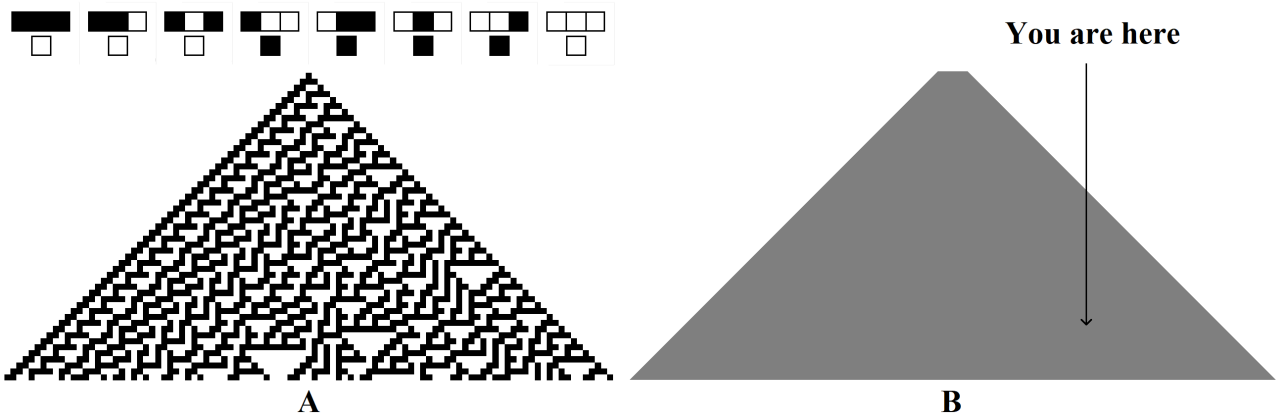

Figure 3: **Your entire life is somewhere in this tapestry**. **A. Rule** 110**.** At the top is the rule to get the color of a cell from the colors of the neighboring cells in the previous lines. At the bottom is the result of applying the rule for 50 iterations, generated in *Wolfram Mathematica*. **B.** A Rule 100 cellular automaton which starts in a particular initial state, chosen to simulate a human person and its environment, including the physical correlates of all the emotions and hopes of that human. Unfortunately the human can't be seen because I had to zoom out too much ☺. But see Theorem 9.

This leads to the following problem:

Problem 1. *If we think that the Rule* 110 *cellular automaton can't be sentient, but a machine can, then what could be the ingredient that the machine has, and the dynamical system from Fig. 3 doesn't? Could it be matter? Could it be the presentist notion of time? Both are metaphysical or ontological assumptions of our minds, with no support from science.*

4.2 Metaphysical baggage

At this point, the reader may feel that, by reducing physics, and in general science, to a mathematical structure, I am reducing it to a strawman, in a desperate attempt to make it too limited to explain consciousness. Physical theories seem much more concrete and down to earth than mathematical structures. So maybe it's the right time for the reader to take a break and try to find some counterexamples of physics that can't be modeled mathematically. There are counterexamples, indeed – theories that are not mature enough to have a rigorous mathematical model, like quantum field theory. Or one can point out apparent conflicts between quantum theory and general relativity, which are again due to our incomplete understanding.

We can find more evident examples by moving to less exact sciences, like biology or neuroscience. Such fields of science may contain explanations that work in some cases, but are false in other cases, leading to contradictions. But the reader should take into account that for sciences

whose object consists of more complex structures, the mathematical formulation becomes more difficult. This doesn't contradict Principle 3, it just shows that some theories we created, as limited human beings, are not complete or mature enough. This should not be seen as a way to avoid the cold conclusion that what they say can be said by a mathematical structure. On the contrary, it just means that they are just approximations of better theories, which can be modeled as mathematical structures at least in principle, by a sufficiently powerful intellect. Even if the reader will appeal to Gödel incompleteness (Gödel, 1931) or Turing's noncomputability result (Turing, 1937), Principle 3 still can hold to make possible such a mathematical model.

A reason for our impression that sciences, including physics, are more than the mathematics modeling various relations, is that most research fields didn't reach enough maturity to eliminate the metaphysical assumptions. In fact, such assumptions are present in various degrees in all theories. Perhaps the best known example in physics is quantum mechanics, for which there are many different interpretations.

But whatever assumptions and concepts we add to our mathematical descriptions of the world, they are in fact drafts, collections of propositions, hence they can be modeled mathematically as well. When we don't have a full mathematical description, we have piecewise descriptions, which may be inconsistent with one another when they overlap, like quantum theory and general relativity, but since the world should be logically self-consistent as a prerequisite of its existence, Principle 3 applies.

The metaphysical baggage is useful, because scientists need some concrete representations of the abstract ideas, and because it is easier to communicate ideas in a less metaphysically free language. But, ultimately, the metaphysical assumptions are not testable, so they are not part of science, unless at some point they are promoted to testable hypotheses.

One may object that, even if a physical theory can be modeled by a mathematical structure, there is something more: we need to map objects and phenomena that we observe in the world to objects in the mathematical structure which models the theory. This is true, but if our theory is rich enough, it should be able to also model the relations between the observers and the things in the world which make the object of the theory.

An important part of the research articles, including of physics, consist of worded explanations, not only mathematical descriptions. Formalizing every argument would make both writing and reading incredibly tedious. The downside is that this facilitates the introduction of hidden assumptions without empirical support.

Definition 16. *We call* metaphysical baggage *those assumptions that, while may serve at building an intuitive representation of the world, have no empirical consequences whatsoever.*

Example 15. *Examples of metaphysical baggage include Newton's assumptions that space and time are absolute, ether theory, and interpretations of quantum mechanics. It does not follow that metaphysical assumptions are necessarily false, some of them can become later part of science. For example, it is possible for an interpretation of quantum mechanics to turn out to have empirical consequences, rather than just reproducing those of quantum mechanics, and become a theory. Other examples are related to time (§3.2) and matter (§3.3).*

4.3 Projection

One may have the impression that if the pattern (Fig. 3) would not be static, but dynamically ran by a computer, this would automatically make it alive and sentient. This is the *strong artificial intelligence* (AI) thesis. However, the same analysis made in the proof of Theorem 6, §3.2 about time and presentism applies here too, so this would make no difference. Taking into account the hardware adds nothing qualitatively new, because what we get is again a dynamical system (see Example 13). The reason why I proved Theorem 9 was to show that the impression that computer simulation or AI can bring something new is not scientifically justified.

Computers store internally the data an process information in a completely different form than that which they present to us on the display. So it would be no difference to the user if internally the computer stores and processes information like the Rule 110 cellular automaton, and displays to us a decoded simulation of a human being demonstrating conscious behavior. In fact, the binary form in which data is stored in our current computers is quite similar to this automaton. The only reason why the 2-dimensional timeless pattern seems to us different from an AI is just that it is encoded in a different representation than the way we see the world. In other words, it makes it more difficult to project our intuitions that physical processes can yield sentience.

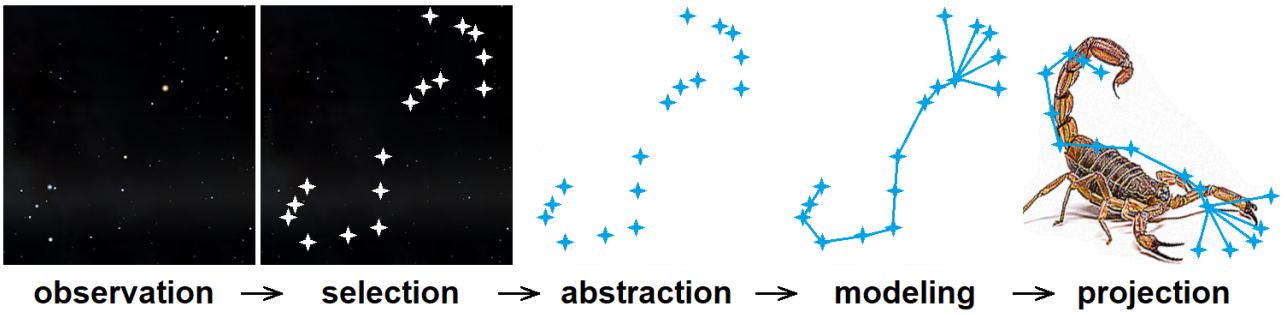

Figure 4: We select some data and ignore the rest, we "connect the dots" to get a model, and we "color" it by projecting from our previous experiences and biases.

We load our descriptions of the world with additional assumptions and intuitions (Fig. 4). This is because the only objective data consists of comparison, hence relations. We have no idea how much they can generalize and exactly what is the generalization, so we guess it, "connecting the dots", and then we fill everything with remnants of intuitions we developed for various situations we encounter. We use the same approach in our mundane life and in scientific discovery. What I did with the Rule 110 example was just to eliminate some unjustified assumptions and intuitions we make. If we stick to just what science can know, which are relations, and avoid adding our intuitions which we use to make science more digestible, this is what we get. Everything else is a projection of our mind. We already saw the example of time, which is present in dynamical systems, but only as a parameter, not as the flowing experience we have. We saw the example of matter, which we think is the reason why science works, but science says nothing about relata, only about relations.

The reason why we may think that sentience reduces to physical processes is our capacity to project emotions. It quite common to have false expectations in love, or to misattribute bad intentions to people who don't even know one's existence. Even for short time, we may love or hate a character from a book or a movie. We can get emotionally attached to virtual pets even as simple as a *Tamagotchi*. Our mirror neurons get triggered because we see signs of consciousness, and not because we actually experience that other beings are sentient.

The argument from §4.1 (Fig. 3) can be seen as an *intuition pump* (Dennett, 2013), although the objective is opposite here: to extract the "air" introduced by our intuition and leave us with the pure relational perspective, the only one that can be called scientific, *cf.* Prop. 1. In particular, it doesn't belong to the class of arguments called by Maudlin "the ploy of funny instantiation" (Maudlin, 1989). Such arguments include *Leibniz's windmill* (section 7 in Leibniz (1989)), Searle's *Chinese room* (Searle, 1980) (with the twist that the funny instantiation is a human bureaucrat oblivious to the computation he or she is performing), time scale modifications (slowing down the time very much) *etc.* Maudlin conceives a theoretical implementation of a Turing machine so that its read/write head does no physical activity, the computation being supported by the combined system of the head and the rest of the machinery. But since

the head is expected to do the same activity regardless of its environment, Maudlin says that this is in contradiction with what he calls the *supervenience thesis*, that "two physical systems engaged in precisely the same physical activity throughout a time will support the same modes of consciousness (if any) through time". Maudlin concludes that his argument implies that *computationalism* (the thesis that consciousness is completely reducible to computation) can't account for experiences, although he thinks that it can support intelligence, understanding, and intensionality (the domain of applicability of definitions). In other words, as phrased later by Chalmers (Chalmers, 1995), that computationalism can only solve the "easy problems" of consciousness. But he thinks that the (physicalist) functionalist "demands that we look beyond the surface behavior to the causal structure of the processes that produce it".

Another argument, based on a more and more "funny" sequence of implementations of a Turing machine, as contraptions made of wood and using sand as a tape, was elaborated by Igor Salom (Salom, 2020), leading to the conclusion that any computation is equivalent to writing a (normally very large) number.

In contrast to the usage of intuition pumps or of "the ploy of funny instantiation", the argument presented in §4.1 goes in the opposite direction, by being derived merely from the observation that science is about relations only, hence from a physicalist/functionalist view on consciousness. The objective was to remove any assumption that our intuition may "smuggle" in the model making it appear to us as conscious, so that instead we can see the naked structure in its full vacuity.

4.4 Representations, interpretations, semantics

Meaning is missing from the timeless 2-dimensional tapestry in Fig. 3 B. The cellular automaton image is just a pattern produced following the eight syntactic rules, and even if it incorporates the full description of the behavior and functionality of a human being, down to the smallest details that can be captured by physics, there seems to be nothing humane there, let alone consciousness. However, if there would be a way to convert the pattern with its syntax into something that has a meaning, this would be a different story.

Semantics is the study of meaning. So maybe one can build a semantic model that would give meaning to the dynamical systems. But whenever we try to do this, what we ultimately do is to find a representation which is just another dynamical system. The only difference may be that we interpret the original dynamical system in terms that seem more concrete to us. Concreteness and intuitiveness are relative to our experience, but they don't contain meaning in the way our sentient experience does.

Let's take the example of natural languages. As children, we are surrounded by people speaking a language. They talk to us, pointing out objects, situations, and actions. We learn to associate the words with the context, and to give them internally a meaning. This meaning is refined and adjusted as the same words are used in different contexts. Our neural networks are adjusted by each new example we encounter, until we fail to notice a relevant difference between the way we use the word, and the way other people appear to us to use it. This process of adjustment and language acquisition may continue for the entire life. But we never reach full certainty that what other people mean by the same word is exactly what we mean. Even when we are talking about apples or rocks or persons, despite being able to identify them, we can't have the certainty that the meanings other people associate to them are the same. We can only know how other people define the words in terms of other words, and how they use them, but not what they mean by them. We may find it difficult to understand what various simple words even mean to us. But we tend to attribute our own meanings to other people using similar words. The meaning is simply not in the language.

There are various theories of semantics. But whenever one tries to develop such a theory, one has the same problems. The first is that it is very difficult to avoid infusing the model

with one's own intuitions and meanings, encoded in our neural networks trained based on the personal experiences. If we manage to get rid of all these decorations, what remains of our theory of semantics are formal relations. Again, relations and nothing more about the relata. Even if intensionality can be realized in such systems, it is devoid of any meaning.

There is no way to build a representation of a dynamical system which is more than a dynamical system. *All representations and interpretations are ultimately just morphisms from the dynamical system we want to represent, to the one which represents it.* Any meaning is put out there by us.

Therefore, I think it is justified to at least consider the possibility that there's something else that gives meaning and allows sentient experience.

5 If sentience is irreducible

There is a crack, a crack in everything. That's how the light gets in.

Leonard Cohen, Anthem

5.1 The "secret sauce"

A major way to understand natural intelligence and consciousness is the investigation of *neural correlates* associated to the processes of consciousness, as in *neuroscience*. Information about the processes of consciousness may be obtained through introspection and reported by the subjects. But introspection is not considered reliable, since the only witness is the subject. Science is based on observations that can be reproduced independently. Thus, it seems more scientific to rely on behavior alone, since this one can be observed independently, while introspection and self-reporting are inherently subjective. The range of what counts as observable behavior extended more and more in time, with the advent of technology able to monitor in more detail the brain activity.

Another approach to study intelligence and consciousness is the field of *artificial intelligence* (AI), and the related, more specialized, *artificial consciousness*. The idea is to reproduce natural intelligent behavior, but also mental processes that are not evident in behavior. So the main position is that of *functionalism*. While AI draws from neuroscience as a source of inspiration, it is not confined to this. A consequence is that it doesn't rely on the biology of the brain. A software is an abstraction of a physical hardware, and it is independent of the micro-states (*cf.* Def. 10). For this reason, the supporters of the *strong AI thesis* (*computationalism*) consider that hardware is irrelevant, so nothing new is gained if the hardware would be biological (Russell and Norvig (2016) ch. 26), and identical function corresponds to identical conscious experience. On the other hand, not everybody thinks that the biological substrate is irrelevant (Searle, 1992). Both neuroscience and AI ultimately study dynamical systems. Neuroscience studies behavior and functionality of brains, so the "hardware" is biological (compare to Example 13). AI studies artificial dynamical systems, abstract or running on hardware, that are able to demonstrate intelligence.

In any attempt to understand consciousness, it would not be justified to dismiss the "easy problems". One should go as far as possible with mapping the neural correlates, and with the physical understanding of the brain architecture and the neurons, of the role played by the biological substrate, and even of the full body, including the *enteric nervous system* (Damasio, 2018). And a great deal of research is going on in neuroscience, evolutionary biology and psychology, artificial intelligence *etc.*

Under these circumstances, is it justified to consider that there is a hard problem? The simplest solution to the hard problem is to explain it away by denying that there is something fundamental about consciousness (Dennett, 1993, 2016; Hofstadter, 2007). Much of what

happens in our minds is indeed illusory, we've seen in §4.3. It may feel natural to think that science will be able to go to the end without breaking, and that there's no need to add a "secret sauce" to explain sentience, and that anything else is an illusion. But shouldn't be someone or something there, to be fooled by the illusions? This is the reason why most supporters of the hard problem consider such explanations to be merely about the easy problems.

I think the arguments I gave so far justify looking at the edge of science too, because this shows that whatever theoretical models we can build, something will be inevitably left outside.

But whatever is that "something" left behind by any theoretical model, it will be left outside here too. Not because the plan to go with Galileo until the end is a limiting decision, but because there is no other way to do it rationally. Anything completely rational, devoid of the intuitions we tend to project by our very nature, is fully equivalent to a mathematical structure. But if science can't say how "the light gets in" to make sentient experience possible, let's use it at least to understand "the crack". So I think it's justified to explore the possibility that there is an irreducible "essence" of consciousness, which I will call *fundamental sentience*. Sometimes it is called derisively *the secret sauce*, metaphor that I find to be quite suggestive, so I will use it, but in a non-dismissive way.

For this, or by curiosity, or to not leave this rock unturned, we will try to see what would happen if the following hypothesis is true:

Hypothesis 1. *Sentience is fundamental.*

But what is sentience?

5.2 What sentience is not

> *The Tao that can be told is not the true Tao; names that can be named are not true names.*
>
> Lao Tzu, Tao Te Ching (Tzu and Minford, 2018).

In Nondefinition 1 I call sentience the ingredient that is supposed to allow conscious experience, and I stated that I don't know to define it. It seems not to be contained in behavior and functionality, even if we go to the best level of detail with the mapping of neural correlates. It is not about "first-person experience" or "self" or "I", because the cases of dissociative identity disorder and split-brain seem to indicate the things are more complicated, and because there are states of consciousness where the boundary between the subject and other subjects seems to vanish (Harris, 2014). I don't think it's the same as *qualia*, the flavor of our experiences, but rather about what makes qualia possible. All of these are incredibly relevant to consciousness, but they are rather manifestations of sentience. Maybe sentience is about *being* (Descartes' *cogito ergo sum*), or that "there is something like to be you" (Nagel, 1974; Chalmers, 2003), in it's essence. As an analogy, one can think what we understand by "matter" as the essence or ontology of the physical objects in materialism. Perhaps a good way to look at sentience, which I will use in the following, is this:

Nondefinition 2. Sentience *is the ontology of experience.*

Claims about ontology are of course metaphysical assumptions, and the best we can know scientifically are its physical correlates (*cf.* §3.4). The meaning of Nondefinition 2 is that I will do the same about sentience, I'll look at the experiential and physical correlates of sentience candidates (EPCSC). Perhaps this can be understood as a weaker version of what Aaronson means by the *pretty-hard problem of consciousness*, defined as "how to construct a theory that tells us which physical systems are conscious and which aren't–giving answers that agree with

'common sense' whenever the latter renders a verdict" (Aaronson, 2014). A "weaker version" though, in the sense that it is limited to identifying physical correlates of sentience candidates, rather than the exact conscious systems, which would require, I think, metaphysical assumptions that seem to run into the sort of problems raised by the *conceivability argument* (Chalmers, 2003).

5.3 Sentience and physical world

Let's consider the collection of all true propositions about all sentient experiences in the world. We don't have to know them, but we can safely assume there exists such a collection of true propositions. It includes propositions about the properties of sentience, both general and at a given time, about the processes by which these properties change, about whatever laws govern these, including consistent descriptions of how sentient beings perceive their sentient experiences. I'll assume they are logically consistent. If these propositions involve other things than sentience, we will include the true propositions about these too. Then, as in the case of physical theories, we will

Corollary 10. *The description of sentient experience admits a mathematical model.*

Proof. It's a direct consequence of Principle 3. □

Remark 4. *The description of sentient experience in Corollary 10 includes consistent descriptions of how sentient beings perceive their sentient experiences. However, something is left outside: is the experience itself, beyond any possible description in terms of logically consistent propositions. This should be understood in relation to Nondefinition 2, that sentience is the ontology of experience. So a "description of sentience" can include a "correlate of ontology candidate" of consciousness, but not the ontology itself.*

Whether or not sentience is a temporal thing, it manifests as something time dependent. So we can strengthen Corollary 10:

Corollary 11. *The description of sentient experience can be represented as a (part of a) dynamical system.* □

We see now that postulating sentience as fundamental is a metaphysical assumption just like postulating matter *cf.* §3.3. There is a difference though, at least to people who think that their experience is evidence for sentience: sentience manifests itself in our sentient experiences. This is not objective evidence, but for those who take it seriously, it is at least subjective evidence. The goal of the following is to see if we can do more than this.

Remark 5. *Whatever sentience is, it must somehow be affected by the world, because it was proposed in the first place as what makes possible for us to experience the world.*

As a consequence of Remark 5,

Theorem 12. *Consider the two dynamical systems, $\mathcal{P} = (S_\mathcal{P}, T, \tau_\mathcal{P})$ representing the physical world, and $\mathcal{S} = (S_\mathcal{S}, T, \tau_\mathcal{S})$ representing the features of sentience from Corollary 11. Then, $\mathcal{S}$ and $\mathcal{P}$ are both completely representable as parts of a dynamical system $\mathcal{W} = (S_\mathcal{W}, T, \tau_\mathcal{W})$, called the* total world.

Proof. The systems $\mathcal{S}$ and $\mathcal{P}$ can be combined, at least trivially, into a a dynamical system $\mathcal{W} = (S_\mathcal{W}, T, \tau_\mathcal{W})$, where $S_\mathcal{W} \subseteq S_\mathcal{S} \times S_\mathcal{P}$, and $\tau_\mathcal{W}(t, (s_\mathcal{S}, s_\mathcal{P})) = (\tau_\mathcal{S}(t, s_\mathcal{S}), \tau_\mathcal{P}(t, s_\mathcal{P})) \in S_\mathcal{W}$ for all $(s_\mathcal{S}, s_\mathcal{P}) \in S_\mathcal{W}$. But since they interact (Remark 5), the combination is not trivial, because the evolution law $\tau_\mathcal{W}$ should include the interaction. The general form of the evolution law

is $\tau_{\mathcal{W}}(t,(s_{\mathcal{S}},s_{\mathcal{P}})) = (\tau_{\mathcal{S}}(t,s_{\mathcal{S}},s_{\mathcal{P}}),\tau_{\mathcal{P}}(t,s_{\mathcal{S}},s_{\mathcal{P}}))$, which includes the possibility that sentience affects the physical world as well.

It is possible that $\mathcal{P}$ and $\mathcal{S}$ are not completely disjoint systems, in other words, that representing the total state by $(s_{\mathcal{S}},s_{\mathcal{P}})$ is redundant, since they may have common parts. In this case, the product $S_{\mathcal{S}}\times S_{\mathcal{P}}$ should be constrained. This justifies assuming only that $S_{\mathcal{W}}\subseteq S_{\mathcal{S}}\times S_{\mathcal{P}}$. In particular, it is possible that the states from $S_{\mathcal{S}}$ are subobjects of those from $S_{\mathcal{P}}$, so $S_{\mathcal{W}}$ can be a curve in $S_{\mathcal{S}}\times S_{\mathcal{P}}$ – the graphic of a function $S_{\mathcal{S}}\mapsto S_{\mathcal{P}}$, so $S_{\mathcal{W}}$ can be identified with $S_{\mathcal{S}}$, or it can be the opposite case, obtained by interchanging $\mathcal{S}$ with $\mathcal{P}$.

In all cases there is a system $\mathcal{W}$ containing $\mathcal{P}$ and $\mathcal{S}$, in particular, $\mathcal{W}$ can be $\mathcal{P}$ or $\mathcal{S}$. □

This includes the possibility that there are multiple sentient beings which interact with one another at least through the physical world (Remark 4).

There are many ways in which two dynamical systems can combine into a larger one, but here I will consider only some cases that can be related to the current theories in physics and interpretations of quantum mechanics.

5.4 Sentience and quantum theory

If sentience is fundamental and $\mathcal{S}$ interacts with $\mathcal{P}$, does this interaction require that $\mathcal{P}$ has nondeterministic jumps? Is $\mathcal{S}$ separate from $\mathcal{P}$, with which it just interacts, in a *dualist* way? Or is it an irreducible and objectively unobservable feature of all physical systems, as in *panpsychism*? Or is sentience the intrinsic nature of physical systems, as in *neutral monism* (Russell, 1921, 1927; Strawson, 2009; Banks, 2014; Goff, 2019)? Answering such questions leads to different views on sentience and its relation with $\mathcal{P}$. Here I'll consider some of the major possibilities.

Type A. $\mathcal{P}$ is nondeterministic, $\mathcal{S}$ is included in $\mathcal{P}$ as subsystem, or as a subset or equivalence class of its properties, and fundamental sentience coincides with its jumping states, exploiting the nondeterminism of $\mathcal{P}$.

Type A can be related to the most accepted view on quantum mechanics, which is that is nondeterministic. In fact, it was thought to be related to the collapse of the wavefunction from the beginning, due to the role of the observer. The idea is developed in particular by Heisenberg (Heisenberg, 1958), von Neumann (von Neumann, 1955), Wigner (Wigner, 1962), Stapp (Stapp, 2004, 2015), and others (Atmanspacher, 2019). Type A is not quite dualism, since $\mathcal{S}$ is not considered to have parts outside $\mathcal{P}$, but it may be property dualism.

There are different nuances in the interpretation, sentience is seen by some in different relations to the wavefunction collapse.

Case A.1. Sentience is causing the collapse

Example (i). In Copenhagen Interpretation (von Neumann, 1955; Wigner, 1962),

Example (ii). One can turn the Transactional Interpretation into an example, by using the interpretation proposed in (Kastner, 2016). Along with the Transactional Interpretation (Cramer, 1986, 1988; Kastner, 2012), other retrocausal *models are compatible with this case (de Beauregard, 1977; Aharonov and Vaidman, 1991; Cohen and Aharonov, 2016; Sutherland, 2017; Adlam, 2018; Friederich and Evans, 2019; Cohen et al., 2019; Wharton and Argaman, 2019).*

Case A.2. Sentience is associated to or caused by the collapse.

In general, this position may seem incomplete, because collapse itself is additional to the Schrödinger dynamics and violates it. A way in which this idea could make more sense is if collapse is part of the dynamics, like in the GRW interpretation (Ghirardi et al., 1986) (see Example 10) or interpretations based on this. An example is the Diósi–Penrose proposal (Diósi, 1987; Penrose, 1996), developed in connection to consciousness by Penrose and Hameroff. Here is a recent review, including discussions of the counterarguments (Hameroff and Penrose, 2017).

Case A.3. $\mathcal{S}$ = collapse events = $\mathcal{P}$. *Some Copenhagen-like interpretations can be understood as meaning that only the collapse events exist (Wheeler, 1990; Wheeler and Ford, 2000; Rovelli, 1996), and in this case they may correspond to conscious experiences (Salom, 2018, 2020).*

An alternative type of models exploiting the wavefunction collapse is

Type B. $\mathcal{P}$ is nondeterministic, and $\mathcal{S}$ is not included in $\mathcal{P}$, but it acts on $\mathcal{P}$ by exploiting the nondeterminism of $\mathcal{P}$.

At first sight, this idea seems to add useless additional structure, that Type A is more parsimonious, and experiments can't distinguish between the two. On the other hand, there is another class of interpretations, proposed with no relation to consciousness at all, which attempt to explain the apparent collapse by using the so-called *hidden variables*, originating with de Broglie and Bohm's *pilot-wave theory* (de Broglie, 1928; Bohm, 1952). In this class of interpretations, particles are point-like, guided by the wavefunction, and they select the branch of the wavefunction rather than collapsing it. A hidden variable theory of this kind is provided by the pilot-wave theory (Bohm, 1952).

Case B.1. Both $\mathcal{P}$ and $\mathcal{S}$ are nondeterministic.
Example (i): In the pilot-wave theory, the hidden variables may correspond to $\mathcal{S}$, and the conditional wavefunction *to $\mathcal{P}$.*
Example (ii): In the pilot-wave theory, the conditional wavefunction may correspond to $\mathcal{S}$, and the hidden variables to $\mathcal{P}$.

In fact, if $\mathcal{P}$ is the full wavefunction, it evolves deterministically, and the point-particles guided by it also evolve deterministically, but in Case B.1 I assumed that $\mathcal{P}$ is the conditional wavefunction, which "collapses" by being updated after each new particle detection, and with respect to the conditional wavefunction, the point-particles evolve nondeterministically as well.
The idea that consciousness corresponds to the point-particles seems to have been suggested in an unelaborated form by Wiseman, who wrote "There is no conscious experience in the empty branches" (Wiseman, 2019).
Some researchers of the pilot-wave theory think that the wavefunction is not a physical entity, but rather has a law-like character (Dürr et al., 1997; Goldstein and Teufel, 2001; Goldstein and Zanghì, 2013), and the point-particles are the only physical entity. This can be seen as inspiring Example (ii) of Case B.1, but also as suggesting the following:

Case B.2. $\mathcal{P}$ is nondeterministic, $\mathcal{S}$ is deterministic.
Example: In the pilot-wave theory, the universal wavefunction may correspond to $\mathcal{S}$, and the hidden variables to $\mathcal{P}$. In this case, $\mathcal{P}$ is fully nondeterministic by itself, but it is determined by $\mathcal{S}$.

Regarding the pilot-wave theory, one may wonder why not taking the physical world $\mathcal{P}$ as including both the wave and the point-particles. This would make $\mathcal{P} = \mathcal{W}$, and $\mathcal{S}$ becomes the subsystem corresponding to the wave or the point-particles, as in the two cases of Type B, and it would be deterministic. This suggests that nothing is gained by assuming nondeterminism, so it is justified to consider determinism.

Type C. *$\mathcal{P}$ is deterministic, and $\mathcal{S}$ doesn't violate its evolution.*

The most straightforward possibility is the following:

Case C.1. Compatibilist sentience. *$\mathcal{P}$ is deterministic, and $\mathcal{S}$ doesn't violate its evolution, but it is compatible with $\mathcal{P}$.*

There are proposals that it is possible to have a deterministic interpretation based on the Schrödinger dynamics alone for a single world, and that the wavefunction collapse is only apparent (Schulman, 1997; 't Hooft, 2016; Stoica, 2016a, 2017). In (Stoica, 2015a) it was shown though that the price is that the initial conditions have to be very fine-tuned to account for definite outcomes of the measurements without discontinuous collapse. This is in fact the alternative to nonlocality that results from Bell's theorem (Bell, 2004). This apparent retrocausality doesn't allow changing the observed past or signaling back in time, just like quantum nonlocality doesn't allow signals to break the speed of light (and by this to use the relativity of simultaneity to violate causality). It is sometimes presented as implying that there is a conspiracy between the initial conditions and the future experiments. Bell called it *superdeterminism* (Bell, 2004), and 't Hooft considers that it precludes completely free-will ('t Hooft, 2016).

Case C.2. Superdeterminism. *$\mathcal{P}$ is superdeterministic, and $\mathcal{S}$ doesn't exist, or is irrelevant and completely determined by and reducible to $\mathcal{P}$.*

But this apparent retrocausality can be interpreted in a timeless manner, using the block universe view provided by the theory of relativity or other eternalist views about $\mathcal{P}$, but such that $\mathcal{S}$ decides what history of $\mathcal{P}$ from the available ones is realized (Hoefer, 2002; Stoica, 2008b,c; Aaronson, 2013; Stoica, 2015a, 2016a, 2017, 2019). This leads to

Case C.3. Postdeterminism. *$\mathcal{P}$ is deterministic, but $\mathcal{S}$ gets to decide what history of $\mathcal{P}$ to be realized among some available options.*

The available options are among the possible solutions of the Schrödinger equation, but only some solutions satisfying some global consistency condition should be valid. It is plausible that these conditions follow from a property of solutions of partial differential equations with topologically nontrivial properties studied in *sheaf cohomology* (Bredon, 1997). If our world is of this type, which is plausible because the fiber bundle formulation of gauge theory shows that nontrivial topology is present, this would mean that the block world itself is subject to global constraints between the choices of Alice and Bob in the EPR experiment (Einstein et al., 1935), and it can be used to reconcile eternalism with free will (Hoefer, 2002; Stoica, 2008b,c; Aaronson, 2013; Stoica, 2019).

Returning to the pilot-wave theory, if we consider as $\mathcal{P}$ or $\mathcal{S}$ the universal wavefunction instead of the conditional one, Case B.2 leads to another possibility. Since the pilot-wave theory requires, in addition to the point-particles, that the wave decoheres, some researchers think that it is not necessarily true that only one of its branches is real in the virtue of the hidden-variables. After all, the empty branches behave just like the real ones. Hence, some think that the pilot-wave theory doesn't solve anything more than the many-worlds interpretation, and in fact it is identical to this one, except that a single branch is considered special, for being singled-out by the point-particles, which are unobservable as such (Deutsch, 1996; Zeh, 1999; Brown and Wallace, 2005). Whether they are right or not, this relationship between the two interpretations allows the Case B.2 to lead to another possibility:

Type D. Many-minds *or* multi-consciousness.
Example: MWI. The overall system $\mathcal{P}$ is the quantum wavefunction with deterministic evolution governed by the Schrödinger equation, and its decohered branches are what their inhabitants perceive as their macro quasi-classical world. The system $\mathcal{S}$ may correspond to a branch, or it may be a multi-consciousness that coincides with $\mathcal{P}$, but what we perceive as our consciousness is relative to our branch.

When Everett introduced the MWI (Everett, 1957, 1973), he proposed that a mind is like a classical system, by having classical memories, rather than quantum ones. This makes only the

parts of the wavefunction that support such system be observable, by the very systems they support, which explains why we perceive only our own branches. This leads to

Case D.1. In MWI, $\mathcal{P}$ corresponds to the universal wavefunction, and $\mathcal{S}$ to each of its branches.
Single-world-consciousness is reducible to classical computation supported by decoherent branches of the wavefunction, which explains why we observe them as quasi-classical.

Some consider that the MWI doesn't support sentience as fundamental precisely because of this functional definition of minds (Salom, 2020).

If we take the view that there can only be a single world populated with minds, we are led to a variant similar to Case B.1:

Case D.2. In MWI, $\mathcal{S}$ corresponds to a single branch of the universal wavefunction $\mathcal{P}$. *The type of consciousness it supports is single-world-consciousness like a classical system.*
Example (i): Single-mind interpretation of MWI (Albert, 2009). Another variant is that $\mathcal{P}$ is the universal wavefunction in MWI, and $\mathcal{S}$ corresponds to classical gravity or spacetime in a hybrid quantum theory on classical spacetime (Kent, 2018).
Example (ii): The pilot-wave theory in which $\mathcal{P}$ is the universal wavefunction, and $\mathcal{S}$ is the system of point-particles, or the associated conditional wavefunction.

The idea of multi-consciousness was proposed by Zeh (Zeh, 1970) as a refinement of Everett's ideas, and it was coined the *many-minds interpretation* in (Albert and Loewer, 1988). It is not prescribed whether sentience is fundamental or not, but if sentience is fundamental, MWI suggests the following possibility:

Case D.3. In MWI, $\mathcal{S} = \mathcal{P}$ is the universal wavefunction.
The type of consciousness it supports is like a quantum system. The particular branches of the wavefunction correspond to single-world-consciousnesses, our quasi-classical minds, and they are not aware of the quantum mind of which they are branches.

From the point of view of sentience, Case D.3 is similar to Case D.2, but there is no special branch that counts as $\mathcal{S}$, rather the entire wavefunction is $\mathcal{S}$, and each branch may support an instance of a mind. This comes with the interesting consequence that when we, as quasi-classical minds, have a choice that corresponds to an apparent collapse of the wavefunction, we get to choose, or rather we are forced to choose, all available possibilities and have all available experiences, in parallel, in different branches.

Many other possibilities could be available, but I tried to confine the discussion to the ones related to physics, in particular related to the main interpretations of quantum mechanics.

6 Does sentience affect the world?

6.1 The switch argument

Let's make a thought experiment. Suppose a person named Bob needs somebody to love, and buys a robot. One that looks and talks like a human, maybe like a loved one who died. Bob's mirror neurons will take it very seriously, even if he may have some beliefs that it is just a machine without consciousness. Now suppose that the robot comes with two possible factory settings: the manufacturer says that Model I is truly sentient, and that Model II imitates or simulates a human person to the deepest details, including reporting to be sentient, without really being sentient. You may want the Model I if you don't want to have doubts that you love and are loved back by a real person. And you may want the Model II if you don't want it to get emotionally attached and suffer for you. And there is a jumper or switch which switches

between the two modes, so that the Model II differs from Model I in that a single module is disabled – call it $\mathcal{S}$-module. What would the $\mathcal{S}$-module do? How deep can the Model II go with the imitation of human beings, without really experiencing anything? Can it report being hungry, like a virtual pet, or like your phone when needs to be charged? Can it display facial expressions that convey emotion, like an emoji? Can it give complicated rationalizations for its actions? Can it say "I love you" like a pre-recorded doll? If all these signs are the result of a complex algorithm, will this make them real? If this algorithm imitates some processes in the brain that lead to similar behavior, does this make them real? What can the manufacturer put safely in the Model II, and what can be put in the $\mathcal{S}$-module? I don't know about others, but I am simply unable to imagine any human behavior that can't go with the Model II, and at the same time I can't imagine what possibly can the $\mathcal{S}$-module do to support genuine sentient experience. No matter what a physical mechanism do, being it a quantum one, I only see a dynamical system equivalent to the Rule 110 timeless pattern.

6.2 The problem of climbing

To make the point I am about to make clearer, I will use a modified version of Searle's *Chinese room argument* (Searle, 1980). Imagine a bureaucratic institution whose employees are grouped in teams which follow certain rules, so that they exchange documents containing information in a way that simulates how neurons exchange information. Each team has to follow the rules so that they behave like a neuron, and the way they connect to other teams to exchange information should model what a neuron does. Such a bureaucratic organization should be incredibly large, since the number of neurons in the brain, ≈ 86 billions, exceeds over ten folds the number of people alive today on Earth. Suppose that all the documents are in a language none of the employees speaks, say Chinese. The reductionist would be able to conceive that such an organization can have the same kinds of experiences and emotions like a human being.

But what would someone who believes that sentience is fundamental think about this? On the one hand, if humans have the needed "secret sauce", since they are the fundamental units of the organization, it may seem that it can be inferred that this means that the organization itself has the "secret sauce". But at a closer look we realize that the organization doesn't make use of the "secret sauce" at all in producing its behavior. Neither its inputs nor its outputs are connected to the level where the "secret sauce" is supposed to exist. So, for someone who thinks the hard problem is really hard, and that sentience is fundamental, it should make no difference at all whether the organization is made of humans who don't understand Chinese, or of philosophical zombies or robots who just follow protocol and have no understanding or experience by design, like Model II in §6.1. So such a supporter of fundamental sentience can't simply infer that, if sentience is fundamental, it automatically endows with sentience any systems that imitates the human brain.

In particular, the $\mathcal{S}$-module making the difference between Model II and Model I in §6.1 should realize the connection between the level where fundamental sentience resides, and the rest of the machine which provides the functionality and behavior of a conscious being.

Before formulating the lesson we learned from this argument, I'll establish some notations, to be used in the following. Consider the dynamical system $\mathcal{B}$ modeling our brain. Since the brain activity relevant to our thinking and controlling our actions appears to act at the coarse grained level, it is modeled by a dynamical system $\mathcal{M}$ (call it *material mind*) which is reducible to $\mathcal{B}$, *cf.* Def. 10. According to Theorem 8, if $\mathcal{M}$ really supervenes on the coarse grained level, it is equivalent to a Turing machine.

If the materialist position is true, this would be all. But if Hypothesis 1 is true and we can think about sentience because it affects the physical world, this means that the dynamical system $\mathcal{S}$ should affect and be affected by (*cf.* Remark 5) the dynamical system $\mathcal{M}$ representing the mind. This means that $\mathcal{M}$ is not only open with respect to the coarse grained level of the

environment, but also with respect to $\mathcal{S}$. The materialist position identifies $\mathcal{S}$ with all material minds like $\mathcal{M}$ in the world. But the position that Hypothesis 1 is true, which states that $\mathcal{S}$ is fundamental, implies that $\mathcal{M}$ should be able to interact to exchange information or signals with the fundamental dynamical system $\mathcal{S}$. Hence,

Theorem 13. *There must be a way for sentience $\mathcal{S}$ to physically affect the material mind $\mathcal{M}$, and for $\mathcal{M}$ to affect $\mathcal{S}$.*

Proof. If materialism holds, the statement follows because sentience $\mathcal{S}$ is included in the union of all material minds like $\mathcal{M}$.

If Hypothesis 1 is true, then the proof was given in the discussion preceding the Theorem. □

This raises the following problem:

Problem 2 (of Climbing)**.** *How does the two-way interaction between the dynamical system representing sentience $\mathcal{S}$ and the material mind $\mathcal{M}$ take place?*

Recalling the discussion in §5.3 and §5.4, we need not only have a way to combine the systems $\mathcal{P}$ and $\mathcal{S}$, but also to find out how they combine such that $\mathcal{S}$ interacts with $\mathcal{M}$. There have to be causal chains connecting $\mathcal{S}$ and $\mathcal{M}$ in both directions. Without such a connection, it makes no sense to say that we are sentient (*cf.* Remark 5) and that we say we are sentient because we are sentient (*cf.* Theorem 13).

One way to understand the Problem of Climbing 2 is in terms of substrate-dependence.

Definition 17. *We say that a dynamical system $\mathcal{B}$ reducible to another system $\mathcal{A}$ is* substrate-independent *if its evolution law is independent of the micro-states,* cf. *Def. 10, eq. (3). Otherwise we say it is* substrate-dependent*.*

From Def. 17 and Theorem 13 we obtain the following.

Theorem 14. *Let $\mathcal{W}$ be the dynamical system combining physics $\mathcal{P}$ and sentience $\mathcal{S}$, such that $\mathcal{S}$ is fundamental. Let $\mathcal{M}$ be a subsystem of a coarse graining of $\mathcal{W}$. Then, if $\mathcal{M}$ is sentient, it is substrate-dependent.*

Proof. Since $\mathcal{S}$ is at the fundamental level $\mathcal{W}$, and since it has to be causally connected to the states of $\mathcal{M}$ *cf.* Theorem 13, it follows that $\mathcal{M}$ is substrate-dependent. □

Remark 6. *In other words, if $\mathcal{M}$ is sentient, it can't be causally closed with respect to its substrate. Of course, if we take into account the interactions with $\mathcal{S}$, we have to describe consciousness not by using $\mathcal{M}$, but a (lower-level) higher-resolution system $\mathcal{M}'$, like a subsystem of $\mathcal{W}$ for which $\mathcal{M}$ is a coarse graining. In terms of the higher-resolution system $\mathcal{M}'$, what appears as random in $\mathcal{M}$ is explained causally in $\mathcal{M}'$. The lower-level system $\mathcal{M}'$ can be substrate independent according to Def. 17, if it is a subsystem of the fundamental level $\mathcal{W}$.*

Remark 7. *Substrate dependence by itself doesn't guarantee that fundamental sentience (*cf. *Nondefinition 2) is able to climb into $\mathcal{M}$ and endow it with sentient experience. It is possible that the Problem of Climbing 2 remains unsolved, but we may still be able to derive consequences from Hypothesis 1.*

6.3 Free will

The idea that the system $\mathcal{S}$ describing sentient experience affects the physical world $\mathcal{P}$ leads to the problem of *free will*. The most spread conception of free will is *libertarianism*, which requires to have more options available in order to make a choice. This view is the default position of both the supporters and the deniers of free will. The common alternative is *compatibilism*, the

position that free will is compatible with the determinism of $\mathcal{P}$, in the sense that even if your choice is predetermined, the important thing is that you get to make it by using your own (also predetermined) will. Other versions of free will are based on chaos theory, uncomputability, and unpredictability (Aaronson, 2013), which explain the "easy problem" of free will, that you feel free because you don't always know in advance what you will choose.

But it's fair to say that there is a *hard problem of free will* as well, and like sentience, it is hard to define.

According to libertarianism, nondeterminism of $\mathcal{P}$ is needed so that $\mathcal{S}$ can use it to make choices. This possibility corresponds to the Types A and B. But since $\mathcal{S}$ has its own laws (*cf.* Theorem 12), we need to take into account for the possibility that $\mathcal{S}$ itself is nondeterministic (Case B.1) or deterministic (Case B.2).

If $\mathcal{S}$ is nondeterministic too (Case B.1), why would this randomness mean free choice? It would be as free as allowing a Geiger counter decide for you. So maybe $\mathcal{S}$ is maximally self-determined, but this means it is deterministic.

If $\mathcal{S}$ is deterministic, and $\mathcal{P}$ is not (Case B.2), then what's the difference from compatibilism? Can we still call this libertarian free will? A difference would be that the true nature of sentient beings corresponds to $\mathcal{S}$, while all the obstacles to its freedom come from $\mathcal{P}$. In this case, even if $\mathcal{S}$ is deterministic, it is predetermined by its own state, just like in compatibilism, and it can use the nondeterminism of $\mathcal{P}$ to express its predetermined choices just like in compatibilism $\mathcal{S}$ expresses itself by affecting the systems that are nondeterministic by the virtue of being open systems.

On the other hand, if the physical world $\mathcal{P}$ is deterministic (Type C), this would apparently leave no freedom for $\mathcal{S}$, unless there is a correspondence between the dynamics of $\mathcal{S}$ and that of $\mathcal{P}$. In this case, we can adopt a compatibilist position (Case C.1), but we can also adopt a post-deterministic position, in which the freedom of $\mathcal{S}$ is powerful enough to decide in an acausal manner what deterministic history $\mathcal{P}$ has (Case C.3). If the correspondence between $\mathcal{S}$ and $\mathcal{P}$ is complete, then such a freedom would be greater than the libertarian one.

But it is conceivable that an even greater freedom is possible, which I will call *plenitudism*: the freedom to explore all options in parallel (Type D, Case D.3). One may find it difficult to call this freedom, since you are forced to explore all available choices. But it can be viewed as if after the branching, each branch of $\mathcal{S}$ is according to its own new state, so it can be made compatible in the same sense as in compatibilism.

Perhaps the maximal freedom is obtained if $\mathcal{P} = \mathcal{S}$ (Cases A.3, C.3, D.3).

The following problem remains open:

Problem 3. *Is there real freedom? If yes, what case ensure it most, and which one is realized in our world,*

1. Libertarianism *(Types A, B, Case D.2)?*
2. Compatibilism *(Case C.1)?*
3. Postdeterminism *(Case C.3)?*
4. Plenitudism *(Case D.3)?*

6.4 The problem of combining the ontologies

In most of the cases listed in §5.4 there is a problem with the ontology: as if it was not enough that we had one ontology problem (§3.4), now there seem to be two or maybe three of them, one for $\mathcal{P}$, and one for $\mathcal{S}$, unified into the single ontology of $\mathcal{W}$.

Problem 4. *If $\mathcal{P}$ and $\mathcal{S}$ have distinct ontologies, how do they combine into the ontology of $\mathcal{W}$?*

But what sort of ontology do we expect in the case of $\mathcal{S}$ compared to that of $\mathcal{P}$? And what sort of objectively observable phenomena do we expect to be associated to $\mathcal{S}$? Do we expect new objectively observable phenomena and properties, in addition to the ones observed in the physical systems? Do we expect $\mathcal{S}$ to require new types of particles and interactions? Why new types, and what should they have that the observed particles and interactions don't have?

From Corollary 11, we know that $\mathcal{S}$ behaves like a dynamical system, which is what $\mathcal{P}$ does too. We also know that $\mathcal{S}$ should play a role in the brain, in particular in its "software" $\mathcal{M}$, and the brain should also play a role in $\mathcal{S}$, and this is the reason to combine it with $\mathcal{P}$.

The only difference, if any, between $\mathcal{S}$ and $\mathcal{P}$ is that $\mathcal{S}$ has a special ontology, which makes possible sentient experience (Nondefinition 2). And from §3.4 we know that we know nothing about the ontology of $\mathcal{P}$. If we arrange these remarks side by side in two columns, we obtain the Table 1.

	Physical structure	**Sentience**
Behaves as	dynamical system $\mathcal{P}$	dynamical system $\mathcal{S}$
Ontology	not specified (§3.3 & §3.4)	postulated to allow sentient experience (Nondefinition 2)

Table 1: The ontologies of the dynamical systems $\mathcal{P}$ and $\mathcal{S}$ when we combine them into the dynamical system $\mathcal{W}$.

As summarized in Table 1, we know of no functional or behavioral difference between $\mathcal{P}$ and $\mathcal{S}$. On the other hand, $\mathcal{S}$ requires a specific ontology, of sentient experience, while the ontology of $\mathcal{P}$, usually assumed to be "materialism", is unreachable by objective means. The simplest way to combine the two ontology seems to be that $\mathcal{P} = \mathcal{S}$, which also means that $\mathcal{P} = \mathcal{S} = \mathcal{W}$.

6.5 Sentientist monism

If we accept Hypothesis 1, we already accept sentience as the ontology for $\mathcal{S}$. This suggests a most parsimonious solution to the Problem 4, simply identify the systems and their ontologies $\mathcal{P} = \mathcal{W} = \mathcal{S}$. By this, whatever features that make us believe in a material ontology of $\mathcal{P}$ are just features of $\mathcal{S}$, *cf.* Corollaries 10, 11 and Theorem 12, they just seem to us "material" because of the projection described in §4.3.

Let us call this proposal *sentientist monism*.

The kind of monism $\mathcal{P} = \mathcal{S}$ is naturally associated to the Cases A.3, C.3, and D.3, but it is conceivable that any sort of physical theory $\mathcal{P}$ can be identified with an $\mathcal{S}$.

Sentientist monism seems to have the following advantages:

1. Simplifies Problem 4, by reducing the ontology of $\mathcal{P}$ to that of $\mathcal{S}$, and making the problem of combining the two into $\mathcal{W}$ evaporate.

2. Because $\mathcal{P} = \mathcal{S}$, it gives the best possible freedom, *cf.* §6.3.

If the hypothesis $\mathcal{S} = \mathcal{P}$ is true and it gives the best possible freedom, this applies to the fundamental level $\mathcal{S}$, but not necessarily to the mind level $\mathcal{M}$, which is a subsystem of the coarse graining of $\mathcal{S}$, even if it can access the fundamental level $\mathcal{S}$. At the fundamental level everything happens physically according to the nature and tendencies of $\mathcal{S}$, because $\mathcal{P} = \mathcal{S}$, but our mundane experiences take place at a coarse grained, macro level, where systems have

impermanent and imperfect features, and seem separate from the rest of the world, which is both a source of resources and of dangers. This means that the maximal freedom at the fundamental level $\mathcal{S}$ may be mostly irrelevant to $\mathcal{M}$, or it is relevant to the extent to which the experience of freedom of $\mathcal{S}$ can climb into $\mathcal{M}$ (*cf.* §6.2).

7 Empirical consequences

So far, I made the case that it is important to at least consider Hypothesis 1, and that it can be explored indirectly by the very means of science and reason. The exploration is indirect because of the inaccessibility of the ontology with the methods of the objective science. But can we derive independently verifiable empirical predictions from Hypothesis 1? I'll argue that, with sufficiently advanced technology, this may be possible. But first, let's explore weaker but more accessible evidence.

7.1 Testing the instrument

A part of this essay consisted in bringing arguments that science can only discover, postulate, and test empirically relations, and everything else is added on top of the cold mathematical structures by our minds. This includes attributing sentience to automata or software, as well as making metaphysical or ontological assumptions about time, matter, consciousness, and artificial intelligence. While more self-aware people or those more inclined to introspection may find this self-evident, it would be nevertheless better if we can test it. Topics like consciousness and artificial intelligence are very serious and important, having tremendous possible impact on our future as humans (Stoica, 2014, 2016b). It would be hazardous to allow our research to be contaminated by our preconceptions and biases, in either direction, but this is hard to avoid. An assessment of the most important instrument of investigation, our mind, is essential.

A way to do this is to monitor the brain activity during the time when researchers think about various problems.

Experiment 1. *Take a cooking recipe, or another algorithm that represents the steps a person can do to obtain an intended result. Put it in a form in which it is evident that there is a person performing all these steps. Then, express the algorithm in pseudocode, in two forms:*

1. *One form of the pseudocode, the* concrete *one, uses descriptive names for the variables and constants, in a way that makes it clear that it is a recipe or another set of instructions for someone to perform. Make it clear what it is about just by using the suggestive names for variables, constants, and procedures.*

2. *The second form, the* abstract *one, is purely syntactical, and all of the names of procedures, variables and constants used in the algorithm are chosen to have no meaning.*

Perform randomized controlled trials in which half of the subjects received a version, and the other half the other version of the pseudocode. Ask the subjects to follow the algorithm, and test their understanding by asking questions, in the same way students are tested in programming classes. In this time, monitor the subjects' brain activity, in particular of the mirror neurons. See if there are any differences, both in the understanding of the algorithm, and in the brain activity patterns. Then, give to each of the subjects the other version of the pseudocode, and repeat.

Probably it will turn out that the mirror neurons are more active when we understand the algorithm as being about people following a recipe, and that this will be associated with the concrete version of the pseudocode. After moving from one version of the algorithm to another,

it is expected that the subjects who read first the concrete one will project the same to the abstract one, and this will be seen in the activity of their mirror neurons.

I think for the most of us the results of this experiments are predictable, assuming it is properly done so that one can indeed monitor the differences. But those who think that this is irrelevant and implies nothing about their views can refine and adapt the experiment to a form and level of detail that they consider relevant.

Experiment 2. *Make similar monitoring and testing as in Experiment 1, this time*

1. *The first group watches a simple android, controlled by simple algorithms and not an AI, performing simple activities similar to some human physical activities.*
2. *The second group watches the software of the android doing these tasks, represented as pseudocode or logical scheme, in a debug mode which is run automatically, highlighting step by step the current pseudocode line or the current step in the logical scheme.*
3. *The third group watches on a computer screen, in one window what the second group sees, and in another window what the first group sees.*

Again, we are interested in particular in the activity of the mirror neurons, in order to see how they correlate with each group.

Remark 8. *In both Experiments 1 and 2, find out the correlations between the results and the opinions of each subject about the strong AI thesis, whether or not sentience is fundamental, whether or not they believe that the substrate matters.*

The purpose of these experiments is merely to make us aware of assumed meaning, prejudice, and bias we project on otherwise meaningless, purely syntactic symbols. It is not evidence for Hypothesis 1, but it is useful to understand how much our previous experience contaminates our understanding, in both directions.

7.2 Subjective evidence

The most common argument in favor of Hypothesis 1 is more or less of the following form

Evidence 1. *We know that sentience is real because it's the first-hand thing we experience or know. Everything else, including empirical evidence, observations, and science, are second-hand knowledge for which sentience is a prerequisite. All we experience, real or not real, have in common one thing: we experience them.*

It is not the fact that we experience sentience that makes it real. We can also experience unreal things. But this is the key point: experience. Even if the things we experience are not real, *the experience in itself is real.*

The belief in sentient experience could be a byproduct of evolution. A byproduct, because it doesn't seem necessary to our survival. What is necessary for the survival of a living being is to act in ways that protect that being, but this doesn't require sentient experience, just favorable instincts and programs. Hence, evolutionism doesn't predict sentient experience, while Evidence 1 corroborates the hypothesis that sentience is real.

But nevertheless this is *subjective evidence*, and not every conscious being agrees with it. In fact, many think that science doesn't leave any room for this, and that sentience is just a computation or a physical process. And many think that, to count as science, *Hypothesis 1 should make objectively verifiable predictions, risking falsification*, and they are right.

Before moving toward more objective empirical predictions, let's consider a bit more if we really should discard from start subjective evidence as pseudoscience.

Bertrand Russell gave a famous example of pseudoscientific claim (Russell, 1969):

> *[N]obody can prove that there is not between the Earth and Mars a china teapot revolving in an elliptical orbit, but nobody thinks this sufficiently likely to be taken into account in practice.*

If an astronaut lost in space is the only one who sees the teapot, maybe nobody will believe him when he returns, but he knows it. Yet, this is not enough evidence for the others, and certainly not enough evidence to count as science. Now consider the huge number of people who claim to see their own "teapot" – to experience their own sentience. This hardly qualifies as absence of evidence, or even as anecdotal evidence. And in fact Russell knew the difference, since he took Evidence 1 seriously, which made him recognize that materialism is not enough, so he supported the idea that there is a "neutral substance" which externally is like physical objects, and internally like sentient experience – a form of neutral monism (Russell, 1921, 1927).

From personal point of view, subjective evidence may be enough, and nobody has nothing to prove to others to be allowed to take their own sentience experience seriously. Some may say that, in order to believe that your own sentient experience is not an illusion or just a matter of computation or of functionality, you have the burden to prove it. You don't. If your own subjective experience tells you otherwise, you have no burden of convincing others about your own experience. You only have this burden if you want others to agree. So this goes the other way too: if your introspection tells you that you are merely computation, you are not forced to accept the opposite just because you can't prove your own experience of lacking sentient experience.

Since both the deniers and the supporters of Hypothesis §1 appeal to their own experience, you can go only this far with Evidence 1. This is why, if one wants to introduce fundamental sentience in science, this should be based on independently verifiable predictions, to put the whole idea to test and risk to falsify it.

7.3 Intersubjective verification

There is a problem with subjective evidence and introspection: you are alone in the process, and more prone to mistakes and wishful thinking. An important reason why objectively verifiable evidence is more reliable is that we can count on our peers to doublecheck it, and to help us avoid confirmation bias and other fallacies. So, if we would like to develop subjective methods of scientific inquiry, we should be much more careful, and develop ways to check one another's reported experiences about sentience – *intersubjective verification tools*. As much as possible, of course, because ultimately nobody can know what's in other people's minds.

At first sight, intersubjective verification is foreign to empirical science. But is there such a thing as purely empirical science? Science consists of a theoretical apparatus, which include laws and calculations. No experiment is purely empirical, since experiments test consequences of hypotheses, and the consequences are derived from fundamental principles or laws by using logical inference and mathematical proof. Moreover, the way the apparatus used in the experiments works also has a major theoretical component, which connects the hypothesized laws to be tested with the quantities that appear on the display of the apparatus. There is no purely empirical data.

And both logical inferences and mathematical proofs require a set of skills which is uncommon in people. Scientists have to verify each other's proofs, and this requires certain expertise, which is not objectively visible in a direct way, but it is tested intersubjectively in a *black box* manner (by exams) and approved by other experts in the same areas, who became experts by the same means. Even scientists can't do more than just trust many of the proofs of their peers, if they don't have enough expertise or enough time to go through the proofs step by step. In fact, there are proofs which we accept and yet no human verified them completely: large collaboration proofs like the classification of the finite simple groups, which took many thousands of pages

(Solomon, 2001; Ronan, 2007), or proofs delegated to the computer (Gonthier, 2008; Krantz, 2011). And even if some peers verify a proof, there is always the possibility of mistakes, since even Hermann Weyl, one of the greatest mathematicians, could believe for a second that 57 is a prime number (H. Weyl (2013), p. 168) [2].

Despite calling physics an empirical science, it mainly consists of a large body of theoretical reasoning and calculation which is unverifiable by most people. And it is safe to say that physicists rely in much of their work on proofs that they couldn't verify personally. Moreover, one should also add to this the fact that some important parts of physics still don't have a rigorous mathematical foundation, but other parts crucially depend on them.

Therefore, it would be an exaggeration to say that physics is 100% objectively and independently verifiable by the peers, and this gap increases in time as new discoveries are added to the main body of physics. Moreover, sciences that are claimed to be reducible to physics, like chemistry, biology, and neuroscience, are developed by people who have to rely even more blindly on the results of their fellows physicists and mathematicians.

Under these circumstances, it is fair to acknowledge that in natural sciences intersubjective verification plays a major and essential role, perhaps greater than the objective independent verification by the peers. If we claim that we should reject from science everything that is not independently verifiable evidence, then we should probably reject all sciences. The only way out seems to be to accept intersubjective verification as part of the scientific method.

But what kind of toolkit can be used for intersubjective verification of sentient experience? Perhaps some techniques of introspection and meditation, designed to contain checkpoints along the way, so that we can verify the achieved progress? This would make the process of intersubjective verification more similar to the testing of mathematical acquisitions of students. I'll not try to develop this idea here, but once we develop a toolkit to explore the world of the subjective empirical, we can proceed to investigate. We can try for example to verify the reports of various people who explored it through meditation and other means of reaching altered states of consciousness. Those who prescribe for example recipes of reaching certain states of consciousness that give certain sentient experiences could try to divide them into elementary steps and criteria of realization for each step, so that anyone who tries them can check for themselves and see if these criteria are satisfied. If possible, also verify them by monitoring the neural correlates.

For example, one could try to verify the proposal of sentientist monism (§6.5), based on $\mathcal{S} = \mathcal{P}$. Very informally, we can derive a prediction.

Argument 1. *If living systems and their minds are subsystems of the coarse grained level, they experience impermanence, separation from the rest of the world, imperfection, opposition from the environment, lack of resources, desire, disappointment, suffering etc.*

Argument 2. *At the fundamental level $\mathcal{P}$, since $\mathcal{S} = \mathcal{P}$, everything happens according to the fundamental evolution laws, in harmony, and they are never broken. Hence, if something changes, it is because the physical system $\mathcal{P}$ had this tendency, and by this, sentience $\mathcal{S}$ itself was in a state that "wanted" that change, and since the tendency to change is followed immediately by an infinitesimal change in the right direction, the system $\mathcal{S}$ realizes its tendencies permanently. The fundamental sentience itself should therefore be in a peaceful state, or in the permanent bliss of continuous accomplishment of its own will, despite the fact that the mind level $\mathcal{M}$ is limited and can be in a state of suffering, unfulfilled desire, fear, worry etc.*

Prediction 1. *From Evidence 1, even if consciousness normally experiences scarcity and suffering (Argument 1), it should also be able to experience the peaceful or blissful experience of the fundamental level $\mathcal{S}$ (Argument 2).*

[2] The same claim is attributed to Alexandre Grothendieck (Jackson, 2004), and for this reason 57 is called *Grothendieck's prime*.

Argument. Consciousness is not completely confined to the coarse grained level, since Evidence 1 says that it has access to the experience at the fundamental level $\mathcal{S}$. Therefore, the experience of the fundamental sentience, which may be freedom, peace, or bliss, should be able to climb through consciousness, allowing it to have these experiences. □

Evidence 2. *Some people claim to have experienced a fundamental state of sentience in which they feel freed, peaceful, blissful, and an experience of unity with everything, beyond any mundane understanding of these words. For an investigation of such states, including subjective experimentation, made by a neuroscientist atheist, see Sam Harris (Harris, 2014).*

Is it possible to turn the above argument and the evidence coming from the testimonies of the few into a more acceptable evidence for a rigorous scientist?

Experiment 3. *Perhaps one can construct a procedure which can be followed in an unambiguous way, and which can be verified independently in some major checkpoints, so that everyone who applied it and masters it, to be able to acknowledge if another person practicing it reached the checkpoints. Additionally, one can verify the reported achievements by verifying their neural correlates. Such a verification of a new person following the procedure by someone who masters it is not very different from the verification of the progress in achieving any other skill, or in learning and understanding a discipline like mathematics for example. If the result is verifiably the one from Prediction 1, it should count as evidence for the proposal that $\mathcal{S} = \mathcal{P}$.*

Despite the significant differences in the form of the methods leading to such states of sentience, which vary from culture to culture and from philosophy to philosophy, there are some common aspects. Examples of cultures and philosophies which explored such states of sentience include Taoism, Advaita Vedānta, Buddhism, Sufism, and even some Judaism or Christian philosophers. In many cases, the practice starts by training the focus, in general by paying attention to the breath in a detached way. Sometimes to a flame, a sound, or a deity. Then, it moves to paying attention to one's own thoughts and experiences, without trying to influence them or to react, just by watching them as they are. This part is very difficult, since our mind tends to wander, any thought triggers more thoughts, and very soon you find yourself following the train of thoughts and forgetting what you intended to do. These processes take place at the coarse grained level, and are in general wanderings of the mind, ruminations, automatisms, programs of the mind. Yet, all these programs start from somewhere, have a cause. For example, one can notice that some thoughts are expressed linearly, phrased in words, and are developed in time like playing a prerecorded phrase, even if the phrase was never used in that form by the person having the thought. Then we can see that even before "playing" the phrased thought, it is already phrased. One can see this by watching how our thoughts unveil, sometimes we can observe that we already have the entire phrase in our mind before starting to "pronounce" it mentally. In time, one can observe how such phrased thoughts originate. One can advance to a level where one knows what will think before thinking, and have the experience of assisting at the birth of every thought we run through our minds. Maybe the phrased thought comes form an image which encodes it, and which suddenly appears in our mind. But similar experience can happen with the mental imagery, since mental images can be watched as they are born in our minds too. In time, we can advance with this process until we are able to prevent thoughts from happening, or to use the infinitesimal split moment when they are born to choose to have different thoughts, or to not have no thought at all, with minimal effort. When we maintain our focus at the root of the thoughts and mental imagery, as we go deeper and deeper, we can have some glimpses or flashes of a fundamental state, at the root of all our mental states, but beyond them, maybe even transcending them. The immersion into this state can be deeper and deeper, until it can be maintained effortlessly. Then, it starts to feel like this was always our natural state, as an absence of separation from other people and the universe, as if all the

separation which we think is natural is in fact an illusion invented by our minds. The state becomes more and more peaceful and blissful, desireless, free.

A possible objection is that similar experiences can be obtained by the use of certain psychoactive substances. Since these are chemicals, one may think that they act at the coarse grained level, which could mean that no climbing can happen from the fundamental level $\mathcal{S}$ to $\mathcal{M}$. But psychoactive substances act by altering or inhibiting certain states of the brain, to produce or allow others, in a way similar to the natural way by which the brain manages itself. So we can't simply reject the possibility on the grounds that these substances operate at the coarse grained level only, without investigating the brain itself to see if it supports the climbing of sentient experience from the fundamental level $\mathcal{S}$ to the coarse grained level of observable and communicable effects. And there are other means to act at the coarse grained level and trigger sentient experiences, *eg.* by sensory input or causing pain. The only conclusion that follows from this counterargument is that sentient experience can be triggered by or associated to physical and chemical means, not that it is reducible to them.

The way to research the relation between $\mathcal{S}$ and $\mathcal{M}$ is by monitoring what happens with $\mathcal{M}$ at the level of physical correlates $\mathcal{P}$ (which is the same as $\mathcal{S}$ in sentientist monism), and see how and where these experiences originate. If they originate at the fundamental level $\mathcal{S}$, one should be able to detect amplifications of phenomena at the fundamental level until they can be differentiated at the coarse grained level. In particular, since the fundamental level is quantum, such amplifications can take the form of decoherence and quantum measurements.

7.4 Objective evidence

The default attitude is that no objectively empirical evidence can support Hypothesis 1, that consciousness is purely a product of the brain, and it doesn't affect the physical world in its turn (*epiphenomenalism*). Even some who consider subjective experience as fundamental prefer to seek refuge in the idea that consciousness only experiences, but doesn't have causal powers.

I think that the following may be the reasons why the default position is that sentience is passive, both in the form that denies fundamental sentience, and in the form that considers sentience fundamental but physically inactive:

1. The objectively empirical data is about physical laws, or reducible to them.
2. There is a temptation to think that if sentience is fundamental and active,
   a) it would have to break down the physical laws, which is impossible by definition,
   b) it should manifest as new physics, *eg.* new particles,
   c) it should manifest in a supernatural way.

But epiphenomenalism is unable to account for Evidence 1, although it is consistent with people reporting to be sentient. If we think that sentience is fundamental because we experience it as fundamental, it has to be causally active, otherwise we wouldn't be able to think and talk about it. Hence, the subjective Evidence 1 becomes objective by having consequences for the physical world.

Corollary 15. *If sentience is fundamental, it should be able to affect the physical world.*

Proof. This is the objectively verifiable part of Theorem 13. From Evidence 1 we know about sentience because it's our first-hand experience. Not only we know, but we are able to think about it, and talk about it, and come up with Evidence 1. Thinking and talking have neural correlates and affects the material mind $\mathcal{M}$, and hence the brain $\mathcal{B}$, and hence the physical world. So, if we take our self-reporting of sentient experiences as being genuinely about sentience and not an illusion, we should accept that our sentient experiences are caused by sentience. □

If $\mathcal{S}$ is fundamental, and acts at the fundamental level of physics $\mathcal{P}$, Corollary 15 leads to:

Prediction 2. *With sufficiently advanced technology, we should be able to identify in the brain a two-way interaction between fundamental physics and the material mind $\mathcal{M}$.*

Prediction 2 runs against the prevalent view that the brain processes relevant for the mind happen exclusively at the coarse grained level.

If $\mathcal{M}$ is a subsystem of the coarse grained level of $\mathcal{P}$, then, to affect $\mathcal{M}$, $\mathcal{S}$ has to affect $\mathcal{P}$. Based on §5, we can identify two main cases consistent with Prediction 2, based on whether the physical system $\mathcal{P}$ is deterministic or not. This makes Prediction 2 branch into two slightly more refined predictions.

Prediction 2a. *With sufficiently advanced technology, if $\mathcal{P}$ is nondeterministic and $\mathcal{S} \neq \mathcal{P}$, we should be able to identify a "glitch" in $\mathcal{P}$, which is used by $\mathcal{S}$ to affect the state of the brain and the material mind $\mathcal{M}$, and which is affected in its turn by them so that $\mathcal{S}$ experiences the processes of the brain.*

Evidence. There is already a "glitch" in physics, if the wavefunction collapse breaks the dynamics governed by the Schrödinger equation. The problem is to find out if this glitch is exploited by the brain, in a way which allows the two-way interaction between the sentience $\mathcal{S}$ and the material mind $\mathcal{M}$. □

A possible solution is *Orch OR* (Hameroff and Penrose, 2017), but there are other ways. In fact, this test will not be able to distinguish among Cases A.1, B.1, B.2, and D.2.

Another possibility is that $\mathcal{P}$ is a coarse graining of $\mathcal{W}$ or even of $\mathcal{S}$. Then, the "glitch" can simply be due to this fact. For example $\mathcal{W}$ may include the wavefunction of the universe and some hidden variables, a variation of Case D.2 Example (ii), with the difference that $\mathcal{P}$ is the coarse grained system obtained by ignoring the hidden variables in the pilot-wave theory. In this case, $\mathcal{S}$ can be the Bohmian point-particles, or even the entire system $\mathcal{W}$, point-particles plus pilot wave.

Experiment 4. *To empirically test Prediction 2a, one should look closely in the brain, to see if there's a place where the wavefunction collapse can introduce randomness in the brain, see if the brain amplifies it to make a difference in its (macro) state. If, in addition, the relevant collapse turns out to be not quite random, but consistent with some declared intention of the brain's user, this would prove Prediction 2a. More precisely, one should be able to identify violations of the Born rule which are essential in the brain's processes. This would require a very advanced technology, but it is possible in principle (Stoica, 2008c, 2012).*

A large scale experiment was reported in (Radin et al., 2016), and analyzed in (Tremblay, 2019), and indicates that small but statistically significant deviations from the Born rule may be present. But, given the large data set required to be analyzed and the possibility of mistakes, this experiment requires more independent verification, either to confirm or to refute the results. Even if it will be confirmed, this will not constitute direct evidence for Prediction 2a, since it is about the wavefunction collapse of a system outside of the brain. If refuted, it will not constitute refutation of Prediction 2a, since the cases making this prediction don't require violations of the Born rule at a distance, outside of the brain.

A similar prediction follows from Case A.3, which, although corresponds to $\mathcal{S} = \mathcal{P}$, is not distinguishable by experiment from the Cases A.1, B.1, B.2, and D.2.

The option of postdeterminism (Case C.3) also leads to a prediction similar to Prediction 2a, but it is slightly different, since it corresponds to $\mathcal{P}$ being deterministic. Since there is no real collapse, there is no "glitch" in the Schrödinger dynamics, but the influence of $\mathcal{S}$ over $\mathcal{P}$ consists in reducing the possible solutions of the Schrödinger to a small subset which matches the allowed outcomes (Stoica, 2015a, 2016a). But since the predictions of quantum mechanics should remain the same, this leads to

Prediction 2b. *With sufficiently advanced technology, if $\mathcal{S} = \mathcal{P}$ is deterministic, we should still be able in principle to test that the resulting state of the system $\mathcal{M}$ interplays with the Born rule in the brain's cells.*

It may seem impossible to distinguish in practice between the Case C.3 and the cases based on collapse or effective collapse (as in the pilot-wave theory or MWI). But in fact there are some consequences of the wavefunction collapse that were not observed experimentally, and if they would be observed, they would falsify the class of single-world unitary interpretations that exemplify the Case C.3. Consider for example Schulman's proposal, presented in (Schulman, 1997) and references therein, which is based on very special states. In (Schulman, 2012, 2016) it is shown that the model predicts that measuring the spin of a particle along a different axis than the one corresponding to the spin of the prepared state exhibits an interaction required to reorient the spin. This is probably very difficult to test experimentally, but if confirmed, approaches based on collapse or effective collapse can be ruled out. Also in (Stoica, 2017) it is shown that a discontinuous collapse, even if just effective, should lead to violations of the conservation laws. This is difficult to verify, due to the fact that the violation is small compared to the same conserved quantities associated to the measurement apparatus, but nevertheless, the claim that there is a discontinuous collapse, even if merely effective, predicts such violations that were never found in experiments, and if they would be found, they would falsify the single-world unitary interpretations.

The Cases D.1 and D.3 may also be tested.

Experiment 5. *In (Päs, 2016), it was proposed that*

> *while the interpretation of quantum mechanics in general and the MWI in particular are notoriously difficult (if not impossible) to test in physics experiments, such tests may be possible in psychology. A simple setup could for example employ probands under the influence of LSD-25 performing quantum measurements (such as spin-up versus spin-down) on a computers screen, while an equally prepared control group deals with an equally looking interface connected to a classical simulation based on a random number generator. It is conceivable that the first group experiences quantum superpositions while the control group does not.*

If this type of experiment works, it can also be used to test the type of fundamental sentience described in the Case D.3, and combined with Prediction 2, it can distinguish it from Case D.1. On the other hand, such experiments should be interpreted with care, because it is most likely that the experience of superposition or coherence is lost once the state of consciousness becomes able to report it, or the neural correlates are measured.

Given that Hypothesis 1 makes independently testable predictions, if we take it seriously, we should probably stop hiding behind epiphenomenalism and risk falsifiability. But even if these predictions will pass the empirical tests, the reader faithful to the tradition of calling everything we discover materialism is free to see this as mere circumstantial evidence, as irrelevant coincidences.

8 Conclusions and open problems

I didn't try to approach the easy problems of consciousness, so all of them are still open problems. I consider the hard problem also open, and possibly it will stay open forever.

What I tried to achieve in this essay was to remove some of the assumptions on which arguments against the idea that sentience is fundamental are based, by using the equivalence between a Turing machine and a two-dimensional timeless pattern. I also brought some counterarguments to some assumptions that are made by some of those who believe that sentience

is fundamental. To avoid making myself such assumptions, I let sentience undefined, but I tried to identify what to expect from its physical correlates. As a consequence, I inventoried or proposed a few possible experiments that can falsify various forms of the hypothesis that sentience is fundamental. They can be performed when the needed technology becomes available.

9 Possible questions and objections

When preparing this essay, I tried to submit it to personal objections, objections usually raised in such discussions, and objections that I anticipate the reader may potentially have. I list some of them that I considered more relevant or likely to be raised.

Objection 1. *Writing about consciousness in a paper is New Age mumbo jumbo.*

Reply 1. *You are welcome to refine this generic objection into more concrete ones, after carefully reading.*

Objection 2. *Still, why trying to connect consciousness with quantum mechanics? Isn't this already criticized for being abused by pseudoscientists? Isn't it known that the brain can't use quantum mechanics at all?*

Reply 2. *There are many ways in which quantum mechanics can be used or abused, so please give specific arguments about this essay, not about other proposals. In this essay, the various interpretations of quantum mechanics are used for the following reasons:*

1. *They provide some concrete examples of relations between the dynamical systems $\mathcal{S}$ and $\mathcal{P}$. Similar examples to those from Types A, B, C and perhaps even D, can be produced with some efforts with specifically designed classical theories alone, but since such classical theories are not related to the known physics and would be forced, I preferred not to do it, and I used existing examples related to our world.*

2. *Physics is quantum, and its various interpretations correspond to various options allowed by no-go theorems like Bell's and Kochen-Specker.*

3. *See some counterexamples to your claim that no quantum effects take place in the review (Hameroff and Penrose, 2017). In fact, the macro level really can't really isolate completely from quantum mechanics, as shown by the violation of the Leggett-Garg inequality,* cf. *Example 10 (Leggett, 2002; Emary et al., 2013; Leggett and Garg, 1985; Leggett, 2008).*

Objection 3. *You are not supposed to quote from Lao Tzu or Leonard Cohen in an article with the pretension to be scientific.*

Reply 3. *I didn't base my arguments or proofs on these quotes. They are there for illustrative purposes.*

Objection 4. *You should leave consciousness to philosophers, because such claims are not testable.*

Reply 4. *I tried to avoid making unjustified claims or hidden assumptions, but if you found some, please let me know. If you read the essay, please remember your objection when reading §7, in particular Russell's argument in §7.2, and the proposed experiments in §7.4. I think at least some of the claims are testable.*

Objection 5. *There is only matter, sentience or consciousness are reducible to it.*

Reply 5. *Do you want to point out if I made some mistakes in §3.3 and §4.1?*

Objection 6. *It is not scientific to even consider the existence of things that can't be proven, like the claim that sentience is fundamental.*

Reply 6. *Good point. But to some of us, sentient experience is the most real thing, and anything else is only possible to be known because of it.*

Objection 7. *One should not speak about things that can't be proven, like the claim that sentience is fundamental.*

Reply 7. *Good point. This is why I took a negative way, to talk about what is speakable only.*

Objection 8. *You didn't define the main object of your essay – sentience.*

Reply 8. *I agree, but you can check Nondefinitions 1 and 2. Also, if you think that I let sentience undefined to smuggle some hidden assumption about it, please let me know exactly where I did this.*

Objection 9. *Your experiments don't prove beyond any doubt that sentience is irreducible or fundamental.*

Reply 9. *No law in science is proven beyond any doubt. The best we can do is to make our theories falsifiable, and to submit them to empirical tests. The experiments I discuss in §7 can falsify some of the claims about sentience, so, if we think that falsifiability is a precondition for a hypotheses to be scientific, here are some possibilities.*

Objection 10. *There are some other possibilities about sentience, which you didn't consider here.*

Reply 10. *I agree. I tried to keep the length of the essay manageable, and to focus on some main possibilities that can be related to the options available in the foundations of quantum mechanics. I am interested in the other possibilities you mention, so thank you for mentioning them.*

Objection 11. *You made some logical or mathematical mistakes, here is where and what these mistakes are.*

Reply 11. *Thank you for pointing them out to me!*

Objection 12. *Here are some places where you used improperly some philosophical terms.*

Reply 12. *Thank you for pointing them out to me!*

Objection 13. *I would like to make some suggestions.*

Reply 13. *Thank you!*

Objection 14. *I read it carefully, but I am still a materialist/epiphenomenalist/dualist/idealist/etc!*

Reply 14. *You are free to adopt any view makes more sense to you.*

Objection 15. *I have no objection about what you wrote, but I think it is dangerous to say that not all systems that behave like humans are sentient. This will lead to discrimination against the artificial intelligence.*

Reply 15. *I think the opposite is more dangerous (Stoica, 2016b).*

Acknowledgment

I am grateful to Adal Chiriliuc, Liviu Coconu, Mihai Prunescu, Robert Ruxandrescu, Igor Salom, Marko Vojinović, and others, for entertaining and challenging discussions.

Bibliography

S. Aaronson. The Ghost in the Quantum Turing Machine. *"The Once and Future Turing: Computing the World," a collection edited by S. Barry Cooper and Andrew Hodges (in press)*, 2013. arXiv:1306.0159.

S. Aaronson. Why I am not an Integrated Information Theorist (or, The Unconscious Expander), 2014. URL https://www.scottaaronson.com/blog/?p=1799.

E. Adlam. Spooky action at a temporal distance. *Entropy*, 20(1), 2018. URL https://www.mdpi.com/1099-4300/20/1/41.

Y. Aharonov and L. Vaidman. Complete description of a quantum system at a given time. *J. Phys. A*, 24:2315, 1991.

D.Z. Albert. *Quantum mechanics and experience*. Harvard University Press, 2009.

D.Z. Albert and B. Loewer. Interpreting the many worlds interpretation. *Synthese*, pages 195–213, 1988.

R. Arnowitt, S. Deser, and C. W. Misner. Republication of: The dynamics of general relativity. *General Relativity and Gravitation*, 40(9):1997–2027, 2008.

H. Atmanspacher. Quantum approaches to consciousness. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition, 2019.

E.C. Banks. *The realistic empiricism of Mach, James, and Russell: Neutral monism reconceived*. Cambridge University Press, 2014.

J.S. Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1(3):195–200, 1964.

J.S. Bell. *Speakable and unspeakable in quantum mechanics: Collected papers on quantum philosophy*. Cambridge University Press, 2004.

D. Bohm. A Suggested Interpretation of Quantum Mechanics in Terms of "Hidden" Variables, I and II. *Phys. Rev.*, 85(2):166–193, 1952.

G.E. Bredon. *Sheaf theory*, volume 170. Springer Verlag, 1997.

H.R. Brown and D. Wallace. Solving the measurement problem: De Broglie–Bohm loses out to Everett. *Found. Phys.*, 35(4):517–540, 2005.

D.J. Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2 (3):200–219, 1995.

D.J. Chalmers. Consciousness and its place in nature. In S.P. Stitch and T.A. Warfield, editors, *The Blackwell Guide to Philosophy of Mind*, pages 102–142. Blackwell Publishing Ltd, Malden, USA, 2003.

C.C. Chang and H.J. Keisler. *Model theory*. Elsevier, 1990.

P.M. Churchland. Eliminative materialism and propositional attitudes. *The Journal of Philosophy*, 78(2):67–90, 1981.

E. Cohen and Y. Aharonov. Quantum to classical transitions via weak measurements and post-selection. In *Quantum Structural Studies: Classical Emergence from the Quantum Level*. World Scientific Publishing Co., 2016. arXiv:1602.05083.

E. Cohen, M. Cortês, A.C. Elitzur, and L. Smolin. Realism and causality I: Pilot wave and retrocausal models as possible facilitators. *Preprint arXiv:1902.05108*, 2019.

M. Cook. Universality in elementary cellular automata. *Complex systems*, 15(1):1–40, 2004.

J.G. Cramer. The transactional interpretation of quantum mechanics. *Rev. Mod. Phys.*, 58(3): 647, 1986.

J.G. Cramer. An overview of the transactional interpretation of quantum mechanics. *Int. J. Theor. Phys.*, 27(2):227–236, 1988.

A. Damasio. *The strange order of things: Life, feeling, and the making of cultures*. Pantheon Books, New York, 2018.

O.C. de Beauregard. Time symmetry and the Einstein paradox. *Il Nuovo Cimento B (1971-1996)*, 42(1):41–64, 1977.

L. de Broglie. *La Nouvelle Dynamique des Quanta, in Solvay Conference, 1928, Électrons et Photons: Rapports et Discussions du cinquième Conseil de Physique tenu à Bruxelles du 24 au 29 Octobre 1927 sous les auspices de l'Institut International Physique Solvay*, volume 24. Paris, Gauthier-Villars, 1928.

D. Dennett. *Consciousness explained*. Penguin UK, 1993.

D. Dennett. *Intuition pumps and other tools for thinking*. WW Norton & Company, New York, 2013.

D. Dennett. Illusionism as the obvious default theory of consciousness. *J Conscious Stud*, 23 (11-12):65–72, 2016.

D. Deutsch. Comment on Lockwood. *The British Journal for the Philosophy of Science*, 47(2): 222–228, 1996.

L. Diósi. A universal master equation for the gravitational violation of quantum mechanics. *Phys. Lett. A*, 120(8):377–381, 1987.

D. Dürr, S. Goldstein, and N. Zanghì. Bohmian mechanics and the meaning of the wave function. In RS Cohen, M Horne, and JJ Stachel, editors, *Experimental Metaphysics: Quantum Mechanical Studies for Abner Shimony, volume 1*, volume 193 of *Boston Studies in the Philosophy and History of Science*, pages 25–38. Boston: Kluwer Academic Publishers, 1997. arXiv:quant-ph/9512031.

A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47(10):777, 1935.

C. Emary, N. Lambert, and F. Nori. Leggett–Garg inequalities. *Reports on Progress in Physics*, 77(1):016001, 2013.

H. Everett. "Relative State" Formulation of Quantum Mechanics. *Rev. Mod. Phys.*, 29(3): 454–462, Jul 1957. doi: 10.1103/RevModPhys.29.454.

H. Everett. The Theory of the Universal Wave Function. In *The Many-Worlds Hypothesis of Quantum Mechanics*, pages 3–137. Princeton University Press, 1973.

S. Friederich and P.W. Evans. Retrocausality in quantum mechanics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019.

G.C. Ghirardi, A. Rimini, and T. Weber. Unified Dynamics of Microscopic and Macroscopic Systems. *Phys. Rev. D*, (34):470–491, 1986.

K. Gödel. Die vollständigkeit der axiome des logischen funktionenkalküls. *Monatshefte für Mathematik und Physik*, 37(1):349–360, 1930.

K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik*, 38(1):173–198, 1931.

P. Goff. *Galileo's error: Foundations for a new science of consciousness*. Pantheon, 2019.

S. Goldstein and S. Teufel. Quantum spacetime without observers: Ontological clarity and the conceptual foundations of quantum gravity. In C. Callender and N. Huggett, editors, *Physics meets Philosophy at the Planck scale*, pages 275–289. Cambridge, Cambridge Univ. Pr., 2001.

S. Goldstein and N. Zanghì. Reality and the role of the wave function in quantum theory. In *The wave function: Essays on the metaphysics of quantum mechanics*, pages 91–109. Oxford University Press, Oxford, 2013.

G. Gonthier. Formal proof – the four-color theorem. *Notices of the AMS*, 55(11):1382–1393, 2008.

G. Grätzer. *Universal algebra*. Springer Science & Business Media, 2008.

P. Pesic H. Weyl. *Levels of Infinity: Selected Writings on Mathematics and Philosophy*. Dover Books on Mathematics. Dover Publications, 2013.

S.R. Hameroff and R. Penrose. Consciousness in the universe – an updated review of the "Orch OR" theory. In *Biophysics of Consciousness: A Foundational Approach*, pages 517–599. World Scientific, 2017.

S. Harris. *Waking up: A guide to spirituality without religion*. Simon and Schuster, New York, 2014.

W. Heisenberg. *Physics and Philosophy: The Revolution in Modern Science*. Harper & Brothers Publishers, New York, 1958.

L. Henkin. The completeness of the first-order functional calculus. *J. Symb. Log.*, 14(3):159–166, 1949.

L. Henkin. Completeness in the theory of types. *J. Symb. Log.*, 15(2):81–91, 1950. doi: 10.2307/2266967.

L. Henkin. The discovery of my completeness proofs. *Bull. Symb. Log.*, 2(2):127–158, 1996.

W. Hodges. *A shorter model theory*. Cambridge University Press, 1 edition, 1997. ISBN 0-521-58713-1,9780521587136.

C. Hoefer. Freedom from the inside out. *Royal Institute of Philosophy Supplement*, 50:201–222, 2002.

D.R. Hofstadter. *I am a strange loop*. Basic books, New York, 2007.

A. Jackson. Comme appelé du néant–As if summoned from the void: The life of Alexandre Grothendieck. *Notices of the AMS*, 51(4), 2004.

R.E. Kastner. *The transactional interpretation of quantum mechanics: the reality of possibility*. Cambridge University Press, 2012.

R.E. Kastner. The Born rule and free will: why libertarian agent-causal free will is not äntiscientific̈. In *Probing the Meaning of Quantum Mechanics: Superpositions, Dynamics, Semantics and Identity*, pages 231–243. World Scientific, 2016.

A. Kent. Semi-quantum gravity and testing gravitational Bell non-locality. *Preprint arXiv:1808.06084*, 2018.

S.G. Krantz. *The proof is in the pudding: The changing nature of mathematical proof*. Springer Science & Business Media, 2011.

J. Ladyman. Structural realism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

A.J. Leggett. Testing the limits of quantum mechanics: motivation, state of play, prospects. *J. Phys. Condens. Matter*, 14(15):R415, 2002.

A.J. Leggett. Realism and the physical world. *Reports on Progress in Physics*, 71(2):022001, 2008.

A.J. Leggett and A. Garg. Quantum mechanics versus macroscopic realism: Is the flux there when nobody looks? *Phys. Rev. Lett.*, 54(9):857–860, 1985.

G.W. Leibniz. The monadology. In *Philosophical papers and letters*, pages 643–653. Springer, 1989.

S. Mac Lane. *Categories for the working mathematician*, volume 5. Springer Verlag, 1998.

T. Maudlin. Computation and consciousness. *The Journal of Philosophy*, 86(8):407–432, 1989.

J. Meiss. Dynamical systems. *Scholarpedia*, 2(2):1629, 2007.

T Nagel. What is it like to be a bat? *Phil. Review*, 83(4):435–450, 1974.

H. Päs. Can the many-worlds-interpretation be probed in psychology? *Preprint arXiv:1609.04878*, 2016.

R. Penrose. On gravity's role in quantum state reduction. *Gen. Relat. Grav.*, 28(5):581–600, 1996.

D. Radin, L. Michel, and A. Delorme. Psychophysical modulation of fringe visibility in a distant double-slit optical system. *Physics Essays*, 29(1):14–22, 2016.

M Rédei and Z Gyenis. Having a look at the bayes blind spot. *Synthese*, pages 1–32, 2019.

M. Ronan. *Symmetry and the monster: one of the greatest quests of mathematics*. Oxford University Press, 2007.

C. Rovelli. Relational quantum mechanics. *Int. J. Theor. Phys.*, 35(8):1637–1678, 1996.

B. Russell. *Analysis of mind*. George Alien and Unwin Ltd; London; The Macmillan Company, New York, London, New York, 1921.

B. Russell. *The analysis of matter*. London: Kegan Paul, 1927.

B. Russell. *Letter to Mr Major*. Allen & Unwin, London, 1969.

S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Pearson Education Limited, Malaysia, 2016.

I. Salom. To the rescue of Copenhagen interpretation. *Preprint arXiv:1809.01746*, 2018.

I. Salom. The hard problem and the measurement problem: a no-go theorem and potential consequences. *Preprint arXiv:2001.03143*, 2020.

L.S. Schulman. *Time's arrows and quantum measurement*. Cambridge University Press, 1997.

L.S. Schulman. Experimental test of the "Special State" theory of quantum measurement. *Entropy*, 14(4):665–686, 2012.

L.S. Schulman. Special states demand a force for the observer. *Found. Phys.*, 46(11):1471–1494, 2016.

J.R. Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

J.R. Searle. *The rediscovery of the mind*. MIT press, Cambridge, Mass., 1992.

R. Solomon. A brief history of the classification of the finite simple groups. *Bull. Amer. Math. Soc*, 38(3):315–352, 2001.

H.P. Stapp. A quantum theory of the mind-brain interface. In *Mind, matter and quantum mechanics*, pages 147–174. Springer, 2004.

H.P. Stapp. A quantum-mechanical theory of the mind-brain connection. In *Beyond Physicalism*, pages 157–193. Rowman & Littlefield Lanham, 2015.

O.C. Stoica. World theory. *PhilSci Archive*, 2008a. philsci-archive:00004355/.

O.C. Stoica. Convergence and free-will. *PhilSci Archive*, 2008b. philsci-archive:00004356/.

O.C. Stoica. Flowing with a Frozen River. *Foundational Questions Institute, "The Nature of Time" essay contest*, 2008c. http://fqxi.org/community/forum/topic/322, last accessed March 23, 2020.

O.C. Stoica. Modern Physics, Determinism, and Free-Will. *Noema, Romanian Committee for the History and Philosophy of Science and Technologies of the Romanian Academy*, XI:431–456, 2012. URL `http://www.noema.crifst.ro/doc/2012_5_01.pdf`. http://www.noema.crifst.ro/doc/2012_5_01.pdf.

O.C. Stoica. The "I" and the robot. *Foundational Questions Institute, "How Should Humanity Steer the Future?" Essay Contest*, 2014. https://fqxi.org/community/forum/topic/2016, last accessed March 23, 2020.

O.C. Stoica. Quantum measurement and initial conditions. *Int. J. Theor. Phys.*, pages 1–15, 2015a. ISSN 0020-7748. doi: 10.1007/s10773-015-2829-2. URL `http://dx.doi.org/10.1007/s10773-015-2829-2`. arXiv:quant-ph/1212.2601.

O.C. Stoica. And the math will set you free. *Foundational Questions Institute, "Trick or Truth: the Mysterious Connection Between Physics and Mathematics" essay contest, third prize*, 2015b. http://fqxi.org/community/forum/topic/2383, last accessed March 23, 2020.

O.C. Stoica. On the wavefunction collapse. *Quanta*, 5(1):19–33, 2016a. http://dx.doi.org/10.12743/quanta.v5i1.40.

O.C. Stoica. Answer to "What are the biggest ways in which the world 20 years from now will probably be different from today?", 2016b. https://www.quora.com/What-are-the-biggest-ways-in-which-the-world-20-years-from-now-will-probably-be-different-from-today-What-are-the-biggest-X-factors-changes-that-are-not-probable-but-are-possible-and-could-be-huge/answer/Cristi-Stoica.

O.C. Stoica. The universe remembers no wavefunction collapse. *Quantum Stud. Math. Found.*, 2017. arXiv:1607.02076.

O.C. Stoica. The post-determined block universe. *Preprint arXiv:1903.07078*, 2019.

G. Strawson. *Mental reality*. MIT Press, Cambridge, Massachusetts; London, England, 2009.

R.I. Sutherland. How retrocausality helps. In *AIP Conference Proceedings*, volume 1841, page 020001. AIP Publishing, 2017.

G. 't Hooft. *The cellular automaton interpretation of quantum mechanics*, volume 185. Springer, 2016.

M. Tegmark. *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Knopf Doubleday Publishing Group, 2014. ISBN 9780307599803.

N. Tremblay. Independent re-analysis of alleged mind-matter interaction in double-slit experimental data. *PloS one*, 14(2), 2019.

A.M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.*, 2(1):230–265, 1937.

A.M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. URL http://www.jstor.org/stable/2251299.

Lao Tzu and J. Minford. *Tao Te Ching (Daodejing): The Tao and the Power, Lao-Tzu (Laozi); Translated with an Introduction and Commentary by John Minford*. Viking, Penguin Random House, New York, 2018.

J. von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1955.

K.B. Wharton and N. Argaman. Bell's theorem and spacetime-based reformulations of quantum mechanics. *Preprint arXiv:1906.04313*, 2019.

J.A. Wheeler. Information, physics, quantum: The search for links. In W.H. Zurek, editor, *Complexity, entropy, and the physics of information*, volume 8, 1990.

J.A. Wheeler and K. Ford. *Geons, black holes and quantum foam: A life in physics*. W.W. Norton & Co., New York, London, 2000.

E.P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Communications on pure and applied mathematics*, 13(1):1–14, 1960.

E.P. Wigner. *Remarks on the mind-body question.* Heinmann, London, 1962.

H. Wiseman. Facebook communication, 2019. https://www.facebook.com/shan.gao.cn/posts/2894260280598948?comment_id=2894277687263874.

H.D. Zeh. On the interpretation of measurement in quantum theory. *Found. Phys.*, 1(1):69–76, 1970.

H.D. Zeh. Why Bohm's quantum theory? *Found. Phys. Lett.*, 12(2):197–200, 1999.

J.K.F. Zöllner. *Transcendental Physics: An Account of Experimental Investigations from the Scientific Treatises of Johann Carl Friedrich Zöllner...* WH Harrison, London, 1880.