On the Psycho-Physical Parallelism

A. Barkai
a.barkai.psyphy@gmail.com

Initially published: December 2018

Abstract
We carefully examine the *psycho-physical parallelism*: the conjecture that one's so-called "subjective experiences" are intimately tied to the state of one's so-called "physical" brain. The parallelism appears obvious when considering rough attributes of one's brain state and of one's conscious experience: one experiences sight when one's brain processes signals from one's eyes, and sound when one's brain processes signals from one's ears. However attempts to *precisely* characterize the psycho-physical parallelism have thus far failed profoundly – giving rise to the conundrum known as the Hard Problem of consciousness.

We find a physics-oriented reformulation of the Hard Problem in the question: *what is the* ***minimal regularity*** *relating "subjective experience" to those structures captured by our existing models of the universe? (structures such as atoms, molecules, biological cells, even brains).* We then trace its essential difficulty to an incompatibility between the structures admitted by our existing models of the universe – and certain qualities we implicitly ascribe to real-world manifestations of "subjective experience".

Considering said qualities of "subjective experience" alongside certain (trivially-observed) couplings between "subjective experience" and effectively-classical physical structures (such as brains), we identify a number of high-level **constraints** which must characterize *any* "consciousness-modeling" structure. From those constraints, we derive 2 high-level properties which must characterize any "consciousness-compatible" universe: effective (i.e. operational) indeterminism, and non-trivial locality violations. Recognizing these 2 qualities in the world described by quantum mechanics (and verified by experiment), we conjecture a non-trivial correspondence between quantum mechanics and the matter of the psycho-physical parallelism. **Hardened skeptics should note that this conjectured correspondence is *not assumed* a-priori, but is rather *derived*.** To preempt misplaced criticism, let us further note that the conjectured correspondence does not suggest a notion of consciousness-induced wavefunction collapse, nor of "free-will"-induced indeterminism.

We proceed to describe a novel emerging physical ontology capturing the verified predictions of known physics (in the form of quantum mechanics) while simultaneously explicitly modeling manifestations of "subjective experience". While incomplete, this ontology can already be seen to offer not only a substantial elucidation of the matter of the psycho-physical parallelism, but also a natural justification of sorts for several uniquely quantum phenomena, including Bell's inequalities violations, the operational indeterminism of quantum systems, the no-communication theorem, quantum teleportation, and superdense coding schemes. **The emerging ontology does not constitute a mere reinterpretation of standard quantum mechanics, as it puts forth a number of detailed falsifiable predictions** regarding the micro-structure of brains (during any given moment). These predictions are rather surprising from the perspective of current-day neuroscience, and thus make for compelling test-cases for the presented proposal.

We must re-iterate that the presented ontology is an emerging one, and is still in its infancy. This paper does not attempt to present a finished-and-done "theory of everything", but rather to recognize and follow a number of crucial and fruitful hints which seem to have been largely overlooked thus far. A number of apparently promising directions for further investigation are also proposed, alongside hints of considerable intersections with other areas of research (such as quantum information science and quantum gravity).

To my wife, who Groks brightly, brightly, and with beauty.

Contents

Introduction

"Any serious consideration of a physical theory must take into account the distinction between the objective reality, which is independent of any theory, and the physical concepts with which the theory operates. These concepts are intended to correspond with the objective reality, and by means of these concepts we picture this reality to ourselves... [In a complete theory] every element of the physical reality must have a counterpart in the physical theory."

— Einstein, Podolsky, Rosen, *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?*

The program of physics seeks to understand the world we inhabit by discovering regularities in observations and experiments. It is the ultimate purpose of this program to distill all found regularities to a minimal set of *irreducible* regularities, dubbed *the laws of physics*. These laws of physics then constitute axioms in a model of the universe which, through logical deduction alone, describes and maximally predicts the results of any observation that could be carried out (if only in principle).

Our aim in this paper is to carefully draw attention to a crucial aspect of this program as described above: it seeks to understand the world *we inhabit*. We do not view the world through a platonic peeping hole, from the outside. Indeed, by definition, our so-called "subjective experiences" are as much a part of the universe as are our nerves, our eyes, and the stars in the night sky.

Nevertheless much tension lies between our current picture of the universe and the self-evident existence of "subjective experience". This tension most clearly manifests when we draw our attention to a perfectly ordinary (if curiously arranged) region of the universe: our own brain.

We typically implicitly accept the *psycho-physical parallelism*: the conjecture that "subjective experience" is intimately tied to the state of a so-called "physical" brain. This parallelism appears obvious when considering rough attributes of one's brain state and of one's conscious experience: one experiences sight when one's brain processes signals from one's eyes, and sound when one's brain processes signals from one's ears. And yet attempts to *precisely* characterize the psycho-physical parallelism have thus far failed completely and utterly, revealing our profound ignorance with respect to the matter.

Our ignorance runs deep – as we are generally unable to conceive of a convincing correspondence between "subjective experience" and *any* structure describable under our current model of the universe. Indeed, even luminaries of modern physics have at times suggested that *no model* of the universe could ever account for the psycho-physical parallelism, rendering it opaque to the endeavors of science[1].

♣ ♣ ♣

Much ink has been spilled on this conundrum, which has come to be known as the Hard Problem of Consciousness. We hope in this paper to meaningfully carry the discussion forward, and lay the groundwork for its migration from the realm of philosophy to that of science (of physics, to be precise).

♣ ♣ ♣

Many of the points raised in this paper lie along axes of thought not typically exercised, and will thus prove *subtle* (though not *complex*). We urge the reader to make a concentrated effort to fully digest the content at every step of the way before proceeding further – paying particularly close attention to text marked in **boldface**. Some sections may prove clearer on a 2nd (or even a 3rd) pass.

This paper is intended to serve not as a definitive treatment of a completed theory, but as an introductory springboard onto a space of still-evolving arguments and ideas. In the interest of clarity, certain trade-offs are made concerning the presentation of key concepts; some sections may prove overly-verbose to some, while other sections tacitly forgo a full treatment of alternative considerations in favor of a simpler route forward. At all times, the intelligibility of the novel core concepts is deemed paramount.

Part I

Philosophical Underpinnings

1 The consciousness instance

"Thought is the thought of thought."

— James Joyce, *Ulysses*

Each of us has a certain notion of that which is referred to by the term "subjective experience". And yet, lacking a model of the universe which accommodates this notion, we are unable to precisely define it.

1.1 A parallel: a working definition for light

A parallel may be drawn between our predicament and pre-Maxwellian attempts to define *light*. According to Wikipedia: *"light is electromagnetic radiation within a certain portion of the electromagnetic spectrum"*. This definition is useful because it reduces an observed phenomenon (light) to a specific manifestation of a *primitive* of our physical theory (the electromagnetic field). However such a reductive definition requires a model of the universe broad enough to accommodate the observed phenomenon. We may ask ourselves, then, how would one go about defining the word "light" *without* possessing a model of the universe which accommodates Maxwell's equations (and by extension, the notion of "electromagnetic radiation")?

The answer is of course obvious to anyone who has successfully learned to associate the word "light" with a certain concept *prior* to studying the physics of Maxwell's equations. A **working definition** for light is simply: *that which is sensed by our eyes*. Though this definition is patently lacking from a modern standpoint, it allows us to begin the search for **the underlying structure** of the observed phenomenon.

As we move onwards, let us bear in mind that even this crude working definition for light requires a model of the universe built on certain primitives. For instance, it builds upon a rough concept of a (3+1)-dimensional universe which is occupied by roughly rigid bodies (such as our eyes), and in which light can travel from one position to another.

Our understanding of each such crude primitive has been radically transformed with the advent of modern physics (ironically, in no small part as a result of our attempts to understand light). Nevertheless each rung in our ladder of conjectured physical primitives held long enough for us to climb closer to the truth.

1.2 *The consciousness instance* and a working definition for "subjective experience"

Let us follow the above example and provide a **working definition** for "subjective experience".

Before we can begin, we must confront the matter of terminology. "Subjective experience" is an inherently deficient term for a structure which we hope to accommodate as an *objective* member of the universe. Since we wish to discuss in this paper *a particular structure* in the realm of that which is often referred to as "consciousness", we shall henceforth use the term **"consciousness instance"** to allude to our object of study. In section (1.3) we will contrast the consciousness instance against related structures which we shall deem the **"consciousness entity"** and the **"conscious entity"**.

We are now ready to state our working definition for the consciousness instance:

The consciousness instance is that which each of us is ***completely*** *sure of.*

We shall briefly draw attention to the structure captured by this working definition so as to bring all readers of this paper up to speed. **Those already familiar with such concepts may safely skip ahead to section (1.3)**.

What is one sure of? A naive first attempt at an answer might go along the lines of *"I am sure that I am sitting at my desk, reading this paper rendered on my computer monitor, having eaten dinner approximately 20 minutes ago"*. One may claim certainty of such information because one perceives it directly, or else has memory of perceiving it directly.

However can one ever be sure that one's perceptions reflect the "real world out there"? We typically assume that our perceptions capture information originating with our bodies' biological sensory organs – which directly measure the "real" environment. And yet our perceptions may just as readily capture information originating elsewhere – for instance with a computer simulation of sorts. This is of course the essence of the philosophical "brain in a vat" argument (as well as of the *Matrix* trilogy): one could never be sure that one's brain isn't in a vat completely isolated from the "real world", being fed sensory information strongly decoupled (even abstracted away) from the underlying substrate of the "real universe".

The naive "brain in a vat" argument makes a clear-cut distinction between the "trusted" brain and the "distrusted" senses. The internal state of the brain – in particular, the *memories* it records – is somehow regarded as more fundamentally trust-worthy than the sensory perceptions presented to the brain. However upon further reflection, we soon come to realize that the veracity of our memories is no less suspect than the veracity of our sensory input; one can never be sure that one's memories were not artificially implanted – recording a past which never was. To call upon our philosophical "brain in a vat" once again, we can imagine futuristic memory-editing surgery being performed on an unsuspecting brain.

A little further along this line of thought, we arrive at a subtle corollary: during any given moment, one cannot be completely sure that one's brain truly stores a lifetime's worth of memories – *real or not.* At any given moment, one is aware of (or: the structure of the consciousness instance associated with one's brain contains) but a thin slice of the totality of one's memories. We all implicitly assume that upon mentally reaching for a specific memory, its content will suddenly come into our awareness – deeming the existence of the *memory* one of those things we are completely sure of. However one can never be *completely* sure that said memory is at all recorded somewhere in one's brain and will indeed come into one's awareness – until it does (to exemplify: think of the name of the place of your birth; the name now occupies your awareness, where only a moment ago it did not).

We may continue to press onwards and ask: given a "brain in a vat", could its associated computer simulation model a universe with laws of physics substantially different from those of *our* universe? The answer, of course, is *of course.* Therefore if *we ourselves* are such "brains in vats", then the physical structure of the "real" universe may be *vastly* different from that of our perceived universe. In such a case *every single one of our assumptions about the universe would be completely without merit* – including the assumption that physical structures resembling brains exist, and that said structures are associated with so-called "subjective experience".

Let us keep this last point in mind lest we be tempted to equate the consciousness instance with the state of a biological brain. Any such attempt would be inherently circular, for the consciousness instance precedes the concept of a brain in our hierarchy of conjecture.

Let us take a step back and remember that the goal of the discussion above is *not* to instill existential doubt in the reader, nor to cast doubt on the existence of structure of the universe which is "merely" conjectured to exist – but rather to *distill the consciousness instance* out of one's mental model of the universe; to draw our attention to that substructure of the universe which happens to be uniquely identifiable through our certainty of its existence.

It may appear as though our working definition for the consciousness instance ("that which each of us is completely sure of") describes an empty structure. That is to say, that one cannot be completely sure of anything at all. However it is not so.

Suppose that one was suddenly presented with convincing evidence that the universe was created last Tuesday at 3:07pm with all of one's memories fully intact, and one is in actuality a $(9\ ^3/_4 + 1)$-dimensional being somehow reminiscent of a giant lizard, stuck in seamless simulation of (3+1)-dimensional planet Earth. Such a revelation would render nearly everything which one *believes* to know about the universe deeply mistaken.

Nonetheless a certain element of one's knowledge would remain unscathed by the revelation: the knowledge of the *experience* associated with such an existence. In other words, even if our experiences relate to the "real world" in unfathomable ways, they nevertheless exist exactly as we imagine them to. As a highly simple example: as one gazes at a drawing of a blue rectangle, **one is certain that the so-called "mental image" of a blue rectangle is a structure which exists in the universe** (even if those structures we call "molecules", "paper", "ink", "eyes", "brains", even "space-time" are not).

Figure 1.1: A blue rectangle, for illustration purposes.

When peeling away every element of our knowledge based on conjecture, that which remains – *experience itself* – is the consciousness instance.

1.3 Terminology: the consciousness instance, the consciousness entity, and the conscious entity

For the sake of clarity, let us explicitly define 2 concepts related to that which we deemed the "consciousness instance".

Each consciousness instance, in a sense, corresponds to *a single time slice* of one's experiences (one is only *completely* sure of one's experience here and now). However the *continuation* of all consciousness instances associated with a single individual over time *also* forms a structure of interest; it is that which we typically designate as the dynamic yet persistent "I" at the core of our being. In this paper, we will refer to such continuations as *consciousness entities*.

Since we typically associate consciousness entities with structures evolving in a (3+1)-dimensional world – people, animals, etc. (much more on this point, in the next chapter) – they too form structures of interest. We will refer to a (3+1)-dimensional structure associated with a consciousness entity as *conscious entity*.

The term "consciousness" is at times used to refer to properties of computationally-complex systems (for example, properties relating to intelligence, or to self-awareness). Note that in this paper, we shall not use the term in this manner.

2 The Physical World and the Hard Problem

"The question which is often asked, whether the world is real or whether we merely dream it, is devoid of all scientific meaning... If our dreams were more regular, more connected, more stable, they would also have more practical importance for us. In our waking hours the relations of the elements to one another are immensely amplified in comparison with what they were in our dreams. We therefore recognize the dream for what it is."

— Ernst Mach, *The Analysis of Sensations*

2.1 Conjecturing the physical world

It is the nature of things that (at least during our waking hours) *tremendous* regularity underlies the structure of each consciousness instance. Having drawn attention to the foundation upon which our mental model of the universe stands, we are now ready to state the obvious: the regularity evident in the structure of the consciousness instance strongly suggests the existence of a universe external to and largely independent from any particular consciousness instance.

Doubtlessly, we could explicitly specify every epistemological steppingstone traversed on the way from an analysis of the consciousness instance to a conjecture of the physical world. Asides from the fact that these steppingstones are normally traversed implicitly and subconsciously (if at all), specifying such a progression is of no use to us here. Let us leave such dubious pleasures to the studious philosopher.

We shall skip straight to the heart of the matter and recognize that humans naturally hold onto a mental model of the physical universe as a roughly Euclidian (3+1)-dimensional arena of sorts, filled with roughly rigid bodies which change their positions over time in roughly similar (and therefore predictable) manners. We further posit that we experience this universe from a particular vantage point in its (3+1)-dimensional spacetime – that of our physical bodies (or more sophisticatedly: that of our brains).

The evidence in support of said external reality is so ubiquitous and overwhelming that in virtually all circumstances, the consciousness instance itself – as an object distinct from the external reality it is *conjectured* to mirror – becomes transparent. We hardly think to ourselves *"I am perceiving sensations best interpreted as the latest instant of time in the life of a 32 year old human, who has a brain packed with certain memories acquired over a lifetime, living in a (3+1) dimensional universe, whose eyes just intercepted light from a lightbulb being turned on"*. Instead, we simply think *"I watched a lightbulb being turned on"*. We typically imagine ourselves to perceive reality directly, as if we indeed had a platonic peeping hole into the universe.

♣ ♣ ♣

With the advent of modern physics, we first formalized and subsequently modified our crude mental model of the universe through careful inspection and experimentation. Today, our fundamental model of the universe as given by the laws of physics is exquisitely backed by experiment in nearly every domain we can probe[1].

2.2 The psycho-physical parallelism

For each consciousness instance, there exists a single conjectured "physical" structure with properties which are directly mirrored in the consciousness instance. This unique substructure of the physical universe is of course one's body – or more particularly, one's brain.

Most naively, one knows that "one's" consciousness instance contains information carried by the photons absorbed by one of two particular collections of particles we call "one's eyes": when one's eyes intercept photons originating with [a drawing of a circle, a drawing of a rectangle, a printout of the binary string

[1] Dark matter and dark energy notwithstanding.

"011001"], "one's" consciousness instance contains [the image of a circle, the image of a rectangle, a representation of the binary string "011001"]. More sophisticatedly, one knows that a collection of particles we dub "one's visual cortex" exhibits similar patterns of activity whenever "one's" consciousness instance bears evidence that one is reading a document. In some cases, one may even know that a much smaller collection of particles we call "a neuron" exhibits similar patterns of activity whenever "one's" consciousness instance carries information related to Jennifer Anniston[2].

2.2.1 The Hard Problem

It is natural to ask: **what is the minimal regularity relating consciousness instances to the state of the conjectured particles of the universe?** This innocent-looking question has thus far resisted every attempt at a resolution, earning itself the name: ***The Hard Problem of Consciousness***.

The *hardness* of the Hard Problem is rooted in the gap between attributes of *structures which can be described by our model of the universe* – and certain attributes which we implicitly ascribe to *"real" consciousness instances which exist in the universe*. Since our model of the universe (and in particular, of the brain) cannot as much as *describe* structures which we would identify as corresponding to consciousness instances – it certainly cannot describe the regularities relating consciousness instances to those structures which *are* expressible under our model (such as particles).

2.2.2 The Easy Problem

To appreciate the Hard Problem, it is instructive to distinguish it from a related problem: what is the mechanism giving rise to the behavior of the (conjectured) collection of particles we refer to as "Bob"? This problem is referred to as the **Easy Problem** in our context.

There isn't much about the Easy Problem which we would normally classify as "easy"; indeed, its resolution lies far outside the reach of mankind's abilities at this point in time. Nevertheless, we call it the Easy Problem because we can easily envision *some* resolution for it, if only far out in the horizon. Particularly, in analogy with a computer, we can imagine a system fully consistent with our current model of the universe which – through complex internal processes – gives rise to all of Bob's observable behaviors.

2.2.3 The Pretty Hard Problem

There exists a variant of the Hard Problem dubbed *The Pretty Hard Problem* by computer scientist Scott Aaronson[3]. This problem concerns itself with the question: *which* physical structures are associated with consciousness instances?

This variant does not require us to objectively characterize consciousness instances in any matter. Nevertheless, it captures the essence of the Hard Problem by demanding of us *some* precise specification of a relation between our existing model of the universe and consciousness instances: structures belonging to domains which we normally hold as somehow orthogonal to one another.

By the end of this paper, we shall have a preliminary answer to this question. Our answer will be quite surprising from the perspective of current-day neuroscience, and will therefore make for a rather compelling testable prediction of our theory.

2.2.4 Aside: on fundamental physics

That the variant of the Hard Problem quoted above is indeed the one we ought to trouble ourselves with is one of the central pillars underlying the ideas presented in this paper. Let us thus briefly justify it through reason and illustration.

Physics may be viewed as the study of systems simple enough to be understood in terms of the irreducible regularities of nature. *Fundamental physics* is therefore, at its heart, simply the study of said irreducible regularities *(and nothing more)*. Any excursion into fundamental physics – into the study of the fundamental structure of reality – ultimately involves 3 steps (which are oftentimes carried-out implicitly, incrementally, and iteratively):

1. The definition of structures constituting the *primitives* of the evolving physical model, and which are to correspond to *elements of physical reality* – to that which *is*.
2. The formulation of *regularities* relating said primitives to one another.
3. The comparison of the model's structure against the structure of observable physical reality[2].

Let us illustrate this notion by partially traversing the implicit chain of reasoning underlying even the crudest of models of reality.

Imagine a person – call him Solo – born with no functional sensory organs. Solo would never evolve a notion of the approximately Euclidean 3-dimensional structure we call "space" (except perhaps as a mathematical curiosity devoid of any meaningful relation to "reality") – for the structure of 3-dimensional Euclidean geometry would not correspond to any of the information available to him. Nevertheless Solo would likely quickly evolve some notion of that which we call "time" – for the idea of continuous linear evolution would capture some of the structure of his experiences. Indeed, Solo would grow to possess a certain *model* of physical reality; a rudimentary and vague model, to be sure, but a model nonetheless. The *primitives* of this model – those structures posited to correspond to fundamental elements of reality – could be approximately regarded as *thoughts*. The *regularities* relating said primitives to one another would be the rough relations between Solo's thoughts at different times.

Though this model of the universe is so crude so as to seem vacant, it reveals itself through those structures which it *does not* admit. Based on his primitive model of the universe, Solo could not conceive of a structure such as a bird, or even a chair as one that could exist in the universe.

Now, imagine further that after many years of existence devoid of sensory experience, Solo's eyesight were somehow suddenly restored. Though Solo would likely find the novel visual information initially incomprehensible, he would soon identify certain illuminating *regularities* in it. For instance, he would notice that there is a certain permanency to substructures of the visual scene – which therefore might as well be regarded as *objects*, and that visual scenes typically shift continuously rather than abruptly (for, unless one is looking directly into a light source which may suddenly change its intensity, a perceived visual scene can change only as one continuously shifts one's gaze in 3D space, or as objects in the scene continuously move-about). Solo would also notice that objects often continuously grow and shrink in size, in a highly correlated fashion (as the objects move closer-to and farther-away from him in 3D space), and even transform in more subtle yet ultimately recurring manners (as the objects rotate about some axis in 3D space). Before long, Solo would piece together such regularities into a coherent picture, evolving a mental model of reality as a (3+1) dimensional Euclidean space-time arena occupied by semi-rigid bodies.

The above portrayal is of course a caricature of sorts of our own experiences. In conjecturing the existence of rigid bodies evolving in a (3+1) dimensional space-time, our own analysis of the structure of physical reality is carried out implicitly rather than explicitly, and indeed, has likely been distilled into our natural cognitive inclinations (which have been shaped over the eons by the processes of evolution). Nevertheless it is important to keep in mind that **we ultimately posit that space-time exists, and that 3-dimensional "particles" are objects which can exist in our universe not because their physical reality is in some way self-evident;** such structures are in no way unique among structures which "could have conceivably" constituted the backdrop of physical reality. **That such mathematical structures capture some of *reality's structure* is fundamentally merely *learned*.**

Now, *having accepted* that the mathematical structure of (3+1) dimensional Euclidean space-time roughly corresponds (in some fashion) to the structure of physical reality – to that which *is*, we further notice that not *any* motion which could *conceivably* manifest in a (3+1) dimensional universe indeed manifests in reality. When kicking a ball, for instance, the ball never travels in a direction perpendicular to the direction of the kick, nor does it ever travel at thousands of km/s relative to the kicker. As we go through life, we grow less and less surprised with the motion of the bodies around us. In other words, we notice that *additional regularity* underlies the structure of physical reality.

[2]Which, incidentally, is captured in the contents of consciousness instances.

One could describe this additional regularity in various ways. One could say: *every time I kick a ball with such and such strength, the ball travels in the direction of the kick with such and such speed; every time solid chunks of some material, of such and such volume ratios*[3] *collide with one other while traveling at such and such velocities, the resulting clamped conglomerate travels with such and such velocity; every time I sit on a stationary cart and throw an object in one direction, the cart begins to move in the opposite direction;* etc., etc... While it is possible to describe the additional regularity underlying reality's structure in this fashion, such a description would be highly redundant. Item 281 in such a delineation would likely be unsurprising given the preceding 280 items. Another way to describe this additional regularity is to say that:

I In an inertial frame of reference, an object either remains at rest or continues to move at a constant velocity, unless acted upon by a force.

II In an inertial reference frame, the vector sum of the forces $\mathbf{F}$ on an object is equal to the mass m of that object multiplied by the acceleration $\mathbf{a}$ of the object: $\mathbf{F} = m\mathbf{a}$.

III When one body exerts a force on a second body, the second body simultaneously exerts a force equal in magnitude and opposite in direction on the first body.

These are of course nothing but Newton's laws of motion[4]. It is sometimes implicitly thought that Newton's laws of motion somehow *explain why* physical processes evolve as they do. They of course do no such thing. Newton's laws merely (axiomatically) *define* a primitive physical framework, as well *describe* the minimal (irreducible) regularity underlying all[5] physical phenomena manifesting under this framework. In the process, they also introduce into our model of physical reality a primitive foreign to the world of Euclidean geometry: that of *mass*.

To drive the point home, let us briefly review 2 additional manifestation of inquiries into fundamental physics.

Here on the surface of the Earth, we notice that objects always fall to the ground when not supported by a force in the upward direction. Once again, we could naively describe this regularity: *when I let go of a rock, the rock falls to the ground; when I let go of a banana, the banana falls to the ground;* etc., etc... We could be more precise: *On the surface of the Earth, unsupported rocks always accelerate in the "down" direction at approximately 9.81 m/s^2; On the surface of the Earth, unsupported bananas always accelerate in the "down" direction at approximately 9.81 m/s^2* etc., etc... We could further distill this regularity: *in the absence of support (or air resistance), all objects on the surface of the Earth accelerate in the "down" direction at approximately 9.81 m/s^2*. If we sent experimenters to the surface of the moon, they would find that a similar regularity holds there, but with a different constant of acceleration: *in the absence of support, objects on the surface of the moon accelerate in the "down" direction at approximately 1.625 m/s^2*.

Seemingly unrelatedly, astronomers likewise notice certain regularities in the motion of the heavenly bodies. For instance, they notice that planets always revolve around stars in orbits approximating ellipses, and that the form of these ellipses is subject to its own regularity, as it relates to the objects' masses, the distances between them, and their relative velocities; no planet has been observed which orbits its star in a triangular orbit. The rate of travel along such an elliptical orbit is itself also subject to regularity; no planet has been observed which arbitrarily speeds up and slows down along its orbit, nor one attaining arbitrary speeds. Once again, we could be more precise and notice, for instance, that when one object orbits a much more massive second object, a line drawn between the objects sweeps equal areas during equal time intervals[6].

Given the framework of Newton's laws of motion, another way to describe *all* of the above regularities (and many more) is simply to say that: $\mathbf{F}_{21} = -G\frac{m_1 m_2}{|\mathbf{r}_{12}|^2}\hat{\mathbf{r}}_{12}$, which is of course Newton's law of universal gravitation. Once again, fundamental physical law does not *explain why* objects fall onto the surface of the Earth, or why the planets revolve around the sun. Instead, it merely *describes* the *minimal (irreducible) regularity* underlying all such phenomena.

[3] I.e. objects of such and such mass ratios.

[4] As quoted from Wikipedia.

[5] The structure of physical reality of course corresponds to the structure of the Newtonian model only approximately, and significantly diverges away from it outside of a limited domain of applicability.

[6] Which is of course Kepler's law of areas.

Finally, let us review another class of regularities found in nature, which (under the framework of Newtonian mechanics) roughly reveals itself thus:

- 3-dimensional objects all belong to 1 of 3 categories, which we may designate as $\{+, -, 0\}$
- Objects in the $(+)$ category experience an attractive force towards objects in the $(-)$ category (and vice-versa, in accordance with Newton's 3rd law of motion).
- Objects in the $(+)$ category experience a repulsive force away from other objects in the $(+)$ category.
- Objects in the $(-)$ category experience a repulsive force away from other objects in the $(-)$ category.
- All objects experience no forces in association with objects in the (0) category.

Upon sufficiently detailed study, many other related phenomena would be revealed, including, of note, phenomena which manifest in cases where the sub-components of the system undergo motion relative to one another (for instance, 2 parallel, electrically neutral current-carrying wires will be attracted to one another when their respective currents flow in the same direction, and repulsed away from one another when their respective currents flow in opposite directions).

Given the framework of Newtonian mechanics, we can distill this entire class of regularities into the form of a few simple equations:

$$\begin{cases} \bullet\ \nabla \cdot \mathbf{E} = \rho \\ \bullet\ \nabla \cdot \mathbf{B} = 0 \\ \bullet\ \nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \\ \bullet\ \nabla \times \mathbf{B} = \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} + \frac{1}{c}\mathbf{J} \\ \bullet\ \mathbf{F}_q = q\left(\mathbf{E} + \frac{\mathbf{v}_q}{c} \times \mathbf{B}\right) \end{cases}$$

These equations are of course Maxwell's equations of electromagnetism. Yet again, fundamental physical law axiomatically introduces new physical primitives into our model of the universe (electric charge, the electric field, and the magnetic field), and again, physical law does not *explain why* physical systems evolve as they do, but rather *describes* the minimal (irreducible) regularities relating the primitives of the model to one another.

Fundamental physics and the Hard Problem

The current-day understanding of the consciousness instance is in a sense analogous to pre-Newtonian conceptions of mechanics. Though we can roughly identify certain regularities governing the relations of consciousness instances to structures captured by our existing model of the universe (such as "particles"), we cannot apply this understanding to arbitrary physical settings, nor can we as much as coherently define the consciousness instance.

We recognize regularities governing the *presence/existence* of consciousness instances. For instance, (philosophical nitpicking aside,) we know that our own brains are always associated with consciousness instances while we are awake. We also recognize regularities governing the *structure/contents* of consciousness instances. For example, we know that at any given time, the contents of "our" current consciousness instances are somehow isomorphic to the information carried by photons which just a moment earlier intercepted our eyes (e.g. when our eyes are fixed on some text, this text manifests in "our" then-current consciousness instance). We also know that whenever we stub our toes on some furniture, the contents of "our" then-current consciousness instances somehow encode the experience of pain. And further, (again, philosophical nitpicking aside) we have good reason to believe that *all* functioning human brains are likewise associated with consciousness instances, governed by regularities similar to those which govern "our own" consciousness instances (more on this, in 2.4).

However given an arbitrary, seemingly brain-like physical conglomerate (such as a modern digital computer, or an alien's biological brain), we cannot with any confidence say whether this conglomerate is somehow associated with consciousness instances *at all*, and if so, what the structure of those consciousness instances might look like.

We are thus led to the definition of the Hard Problem as previously stated: **what is the minimal regularity relating consciousness instances to those structures captured by our existing model of the universe (such as particles)?**

We can perhaps now better appreciate that the essential difficulty of the Hard Problem is rooted with our inability to as much as *describe* structures which we would identify as corresponding to consciousness instances. For the regularities relating consciousness instances to those elements of physical reality we designate as 'particles' cannot be formulated without also defining structures which convincingly correspond to reality's consciousness instances.

2.3 Classically-incompatible qualities of the consciousness instance

As we shall soon argue, 2 of the qualities which we implicitly ascribe to the consciousness instance are incompatible not only with our existing model of the brain, nor even with our existing model of the universe – but with an entire *class of models* which may best be described as *classical models* of the universe. As such, no model belonging to this class could conceivably resolve the Hard Problem (by identifying the regularities relating consciousness instances to the conjectured particles of the universe) – for no such model could as much as accommodate structures modeling consciousness instances.

♣ ♣ ♣

The 2 classically-incompatible qualities of the consciousness instance are its *integration of complex information*, and its *non-isomorphism to classical information*.

2.3.1 Integration of complex information

In classical models (which – to reiterate once again – are assumed to model the brain's cognitive operations well), composite structures are a convenient yet ultimately redundant and incomplete abstraction over a fundamental particulate reality. In certain circumstances, a collection of primitives of a classical model may be efficiently *abstracted* as a single entity, governed by a set of *emergent* higher-level regularities. However such emergent abstractions are merely an *effective description* of the behavior dictated by the fundamental lower-level regularities – and are therefore transparent to the underlying model.

♣ ♣ ♣

Though it is difficult to formally distinguish between *fundamentally integrated* structures and *composite* structures, the intuition for such a distinction is clear: to be regarded as *fundamentally integrated*, the integrity of a structure must be inexorably and uniquely pointed to by the underlying model, rather than by a high-level analysis of the model's emergent properties.

Since a model is nothing but a specification of regularities relating structures to one another, we may reasonably require that **for information to be regarded as fundamentally integrated *with respect to a given model*, it should be impossible to state the regularities of the model in terms of subsets of the information**.

♣ ♣ ♣

A relevant demonstration of this notion can be found in the abstractions underlying the operation of the modern computer.

Modern programming languages allow for the definition of custom structures which play a role in the computer's behavior as described by its simplified behavior model: the program. For instance, when designing a program intended to draw 2-dimensional shapes, a programmer may define a `Shape2D` structure – whose instances are *semantically guaranteed* (within the appropriate level of abstraction) to represent entire pixel bitmaps of a certain dimension. The programmer may even define functions which take such `Shape2D` instances as arguments (e.g. `isARectangle :: Shape2D -> Bool`) – thereby formulating the regularities of the computer's behavior model in terms of `Shape2D` instances (e.g. if `isARectangle` returns `true` when called with the `Shape2D` instance rendered directly under the user's finger, a certain group of pixels turns green). According to our earlier definition, the existence of such functions deems the information encoded

by a single `Shape2D` instance **fundamentally integrated** *with respect to that particular level of abstraction*, for the regularities of the model are stated in terms of such instances.

However when a program is translated from an entity in the platonic realm of semantic structures to a physical state (of a group of particles we call "a computer") *effectively described* by said semantic structure, instances of the `Shape2D` structure are demoted to the status of *effectively integrated structures*, whose effective integration arises out of the emergent behavior of primitives of a more refined model of the computer's behavior.

Since the computer is an engineered system, the `Shape2D` instances described by the high-level programming language are coupled to the physical substrate of our universe by a conveniently clean hierarchy of effective descriptions. First, those functions which were earlier defined in terms of (whole) `Shape2D` instances become realized as a succession of functions defined on sub-structures of `Shape2D` instances ("properties" of `Shape2D` instances). With respect to this more refined model of the computer's behavior, then, the `Shape2D`'s *properties*[7] are those structures deemed fundamentally integrated.

Before long, this more refined description is reduced once again, as *functions* become realized as successions of CPU instructions operating on individual bytes rather than on any of the semantic structures defined through the high-level programming language. At this level of abstraction, then, it is the *byte* which constitutes a unit of fundamentally integrated information.

The CPU instructions are themselves ultimately realized through chained logic gates, operating on merely 1 or 2 bits at a time – deeming such individual bits fundamentally integrated; the logic gates are in turn realized through transistor circuits, which operate not on bits at all but on electric voltages... It is at this point that a more direct relation to our universe's physical substrate ensues, as idealized circuits are realized through the interactions between real atoms and electrons.

At this more descriptive (and more accurate) level of abstraction, the apparent integrity of the information corresponding to a single `Shape2D` instance is revealed to be nothing but a (typically convenient) illusion; the atoms whose properties ultimately encode the bits representing an individual `Shape2D` instance are no more related to one another than to any other atom in their vicinity. Though a treatment of `Shape2D` instances as integrated wholes surely simplifies a high-level assessment of the computer's behavior, such an *effective* abstraction – like all effective abstractions – would ultimately break down if sufficiently prodded, giving way to the richer underlying reality. No amount of source code would accurately capture the behavior of a computer operating in the presence of powerful cosmic radiation – nor of one dropped into a volcano.

According to the modern view of the brain, a similar (though less clean) abstraction chain couples the high-level operations of the brain to the physical substrate of our universe. In particular, we commonly maintain that *effectively-classical primitives* account for the behavior of the brain relevant to cognition.

And yet each consciousness instance harbors complex internal structure which seems to roughly correspond to complex classical information (for instance, as one gazes at a picture of a rectangle, a substructure of "one's" current consciousness instance is somehow isomorphic to the shape of a rectangle). **How could we ever compute the (complex) information encoded in a single consciousness instance from an effectively-classical model of the brain?**

We must again reiterate: we are not speaking here of the account matters according to our *conjectured* model of the universe, in which *conjectured collections of particles* we call "people" (including the "person" we designate as "me") *effectively behave as though* they saw a rectangle in "their" mind's eye. The above could be accounted for in a manner similar to our earlier account of the `Shape2D` instance: as a high-level abstraction which efficiently captures emergent characteristics of the structure dictated by the fundamental low-level regularities. Instead, we are speaking of the consciousness instances alluded to by our working definition ("that which each of us is completely sure of")[8].

Surely, we can come up with any number of dictionaries which map the states of a given classical model to units of complex information. We could consider a `(2m x 2m x 2m)` volume to be a fundamental "consciousness volume" in which particle positions encode the information of a single consciousness instance, or perhaps somehow compare the relative velocities of particles to the digits in the decimal expansion of

[7] Or the properties' properties, ad infinitum.

[8] If you find yourself confused by the distinction, I urge you to read part (I) again.

π. Integrated Information Theory (IIT), which similarly recognizes the tension between the effectively-classical world and the apparent integration of the consciousness instance's information content, identifies a particularly elegant and aesthetically pleasing mapping in the long-term causal relations between primitives of the model[4]. Yet no matter how aesthetically pleasing, all such dictionaries are at their core arbitrary,

for the supposed "integrity" of the complex information in no way affects the model's behavior – and is therefore *undetectable* and *unverifiable*.

♣ ♣ ♣

Another way to appreciate this incompatibility is to ask ourselves: suppose that a model existed which convincingly and reasonably accounted for the information contents of individual consciousness instances; what would such a model have to look like? A classical model would forever leave us scratching our heads, for the identification of any[9] singular unit of complex information in a classical model would necessarily seem arbitrary and unverifiable.

2.3.2 Non-isomorphism to classical information

A characteristic common to all of our existing models of the universe is the isomorphism between the possible states of each model and sequences of classical bits. In other words, all possible states are *fully expressible*, in some manner, using classical information. Indeed, to an extent, the expression of phenomena using transformations over classical information is the very essence of our *idea* of a model.

However, in the case of the consciousness instance, this characteristic is at odds with our experience; we simply cannot imagine how classical information could ever *fully* capture the consciousness instance.

♣ ♣ ♣

To appreciate this, suppose we were presented with a candidate theory purporting to fully resolve the Hard Problem. We can imagine, for instance, that in this theory a physical consciousness instance would somehow correspond to a 42-dimensional (fundamentally integrated) vector. Such a theory would be far from useless; it would make clear predictions with respect to the presence of consciousness instances in a given physical setting (resolving the Pretty Hard Problem), as well as with respect to the information content of each consciousness instance. We might study such a theory in detail, and discover that all of its predictions indeed hold – shedding much-needed light on the place of the consciousness instance in our universe.

However we could never accept such a theory as a final account on the matter of the Hard Problem, for we have an unshakable intuition that a certain aspect of the consciousness instance would remain not at all illuminated by such a specification. Given the 42-dimensional vector describing consciousness instance `X`, we could always ask: *"why does consciousness instance* ***`X`*** *feel like* ***this*** *and not like* ***that****?"*. We would hold that the theory is missing a certain mapping from consciousness-instance-describing structures which exist in the model to consciousness instances which exist in the real universe.

More poetically, we cannot fathom the *experience* of color being described to the blind.

♣ ♣ ♣

It is likely this quality of the consciousness instance which is most often taken to spell doom for our entire endeavor, rendering the consciousness instance forever beyond the grasp of reason. In (4.2), we will show that our predicament is not quite so gloomy, and that even this characteristic of the consciousness instance can be absorbed into a scientific theory.

[9] Other than perhaps the complex information describing *the entire universe* captured by the model – which could not account for the information contents of the disparate consciousness instances manifesting in our universe ("one's" consciousness instance is somehow isomorphic to a rectangle when one's eyes gaze at a rectangle, and to a triangle when one's eyes gaze at a triangle).

2.4 Pinpointing our assumption of the existence of external consciousness instances

Earlier, we described the Easy Problem as arising from the question: *what is the mechanism giving rise to the behavior of the* ***collection of particles*** *we refer to as "Bob"?*

Though we are unable to describe any precise relation between consciousness instances and matter, we implicitly assume that in addition to the Easy Problem, the Hard Problem likewise applies to "Bob"; in other words, that those *particles* referred to as "Bob" are associated with consciousness instances. For the sake of completeness, let us explicitly point out the chain of reasoning implicitly driving us towards such a conjecture:

1. The consciousness instance exists.
2. Much regularity underlies the structure of each consciousness instance, strongly suggesting the existence of a "physical universe" external to the consciousness instance. Therefore, we posit, the "physical universe" exists.
3. "One's" consciousness entity bares a unique correlation to a specific structure in the "physical universe": one's brain.
4. The "physical" universe contains many structures very similar to one's physical brain. We call those structures "other people's brains" (and to an extent, in decreasing order: brains of intelligent mammals, all mammals, animals at large, perhaps even computers). The similarities between one's brain and others' brains extend to behavior, form, and manner of origin.
5. One can find no difference significant enough to explain why one's own physical brain would be associated with a consciousness entity while others' would not not be. Therefore one maintains, without possessing a detailed understanding of the regularities relating consciousness instances to so-called matter, that others' brains are likely also associated with consciousness entities.

3 The Consciousness-"Matter" Interaction

"What is most striking is not so much that Parmenides and Zeno were wrong as that they did not bother to explain why, if motion is impossible, things appear to move. Indeed, none of the early Greeks from Thales to Plato, in either Miletus or Abdera or Elea or Athens, ever took it on themselves to explain in detail how their theories about ultimate reality accounted for the appearances of things. This was not just intellectual laziness. There was a strain of intellectual snobbery among the early Greeks that led them to regard an understanding of appearances as not worth having. This is just one example of an attitude that has blighted much of the history of science... Are we now making similar mistakes, passing up opportunities for scientific progress because we ignore phenomena that seem unworthy of our attention?"

— Steven Weinberg, *To Explain the World: The Discovery of Modern Science*

In chapter (1) we established that the consciousness instance is the only object we can be absolutely sure exists. In chapter (2) we acknowledged the rich reality strongly conjectured to exist based on an analysis of the contents of consciousness instances, as well as the unique correspondence between certain sub-structures of that reality (our bodies, our brains, our neurons) and consciousness instances themselves. Let us continue our effort to explicitly confront that which is normally implicitly assumed. In the process, we will learn something novel of the nature of the consciousness instance.

3.1 Can the consciousness instance be measured?

Deliberations on the Hard Problem usually focus on the classically-indescribable attributes of the consciousness instance, which are taken to be *passive* with respect to the rest of the universe. These passive attributes are described in terms of *qualia*: those building blocks of the consciousness instance which distinguish one instance from another. A favorite example is "the redness of the color red" (which is different from the "blueness of the color blue", or the "C-ness of the note C"). Examples of questions which are commonly asked are *why does the color red* ***feel*** *as it does?*, and *do all people experience the same sensation in association with the word 'red'?*

It is commonly assumed that each brain state is accompanied by a passive subjective experience of certain qualia – and that this is the only meaningful observation which can be made with respect to the consciousness instance.

Were this the case, the consciousness instance would forever remain beyond the reach of science; to scientifically resolve the Hard Problem, or even just the Pretty Hard Problem, the existence of the consciousness instance must be somehow inferred from observations of effectively classical physical structures[1] (such as a lab instrument's dial). One cannot scientifically probe that which is not measurable, nor can one measure that which has no measurable effect.

Thus a scientific theory of the consciousness instance hangs on the question: *can the consciousness instance affect effectively-classical physical matter?*

3.2 A false-start: the appearance of "free will"

A characteristic of the consciousness instance we have thus far refrained from commenting on is the persistent appearance of so-called "free will". We generally associate "ourselves" with consciousness entities, and feel as though "we" can actively control our physical bodies' actions as we please. Asides from the matter that it seems nigh impossible to as much as coherently *define* "free will" into a model of the universe built on unbroken regularities, it is very easy to imagine that the feeling of "having free will" arises out of yet another underlying physical brain state reflected in the consciousness instance as qualia. We can imagine that "having free will" is a (presumably useful though illusionary) abstraction encoded in our brain's synaptic

[1] If we are being pedantic, one only ever measures "one's" consciousness entity. However for all practical purposes, information originating outside of one's skull only enters "one's" consciousness entity through one's senses – which can only convey information about effectively classical structures.

structure, and that this abstraction is reflected in the consciousness instance as the thought "I have free will" – not much differently from the manner in which the synaptic activity associated with seeing the color red is thought to be reflected in the consciousness instance as qualia of perceiving "redness". As such, our intuition of possessing free will is not admissible as evidence of the measurability of the consciousness instance.

3.3 Extracting insight from our knowledge of the consciousness instance

A crucial yet often-overlooked aspect of the consciousness instance comes to light upon asking the simple question: *how has the topic of qualia come to be discussed in philosophical papers?*

For discussions of qualia to appear on paper, ink molecules (for the sake of argument) must be arranged on paper to define words and sentences. Let's imagine tracing the causal chain leading to the placement of one such molecule in its final position.

Immediately prior to appearing on paper, the ink molecule was ejected from a printing head due to a pressure differential, caused by the expansion of superheated ink. The superheating resulted from the conduction of electric current through a material offering significant electric resistance, resulting in the heating of the material, and the subsequent conduction of the heat to nearby ink molecules. The electric current was conducted due to a voltage difference between the 2 ends of the material, and the resulting heat energy was transferred to the ink molecules by way of kinetic diffusion and electromagnetic radiation...

We may continue tracing the causal chain in this fashion, to arbitrary precision, essentially solving our physical model's equations of motion for the entire system. In due course, this causal chain would lead us back to the muscle contractions of the paper-author's fingers typing away at the keyboard, then to the electric signals carried along the author's nerves (triggering the muscle contractions), and eventually, to the author's brain. When we continue tracing the causal chain within the author's brain, we will find 1 of 2 things to be true:

3.3.1 Possibility (I): passive consciousness instances

It is possible (and indeed *expected* by the physics and neuroscience communities) that we will find all of the brain's computational operations to be well-modeled by classical and quasi-classical[2] abstractions.

In such a case, the supposed existence of the consciousness instance (as a partially-classically-indescribable structure containing complex information) would have played *no role whatsoever* in the causal chain concluding with the writing of the paper which claims to discuss the characteristics of the consciousness instance. Indeed, if one accepts that the consciousness instance mirrors in some fashion the state of a "physical" brain, then this possibility would further imply that the existence of the consciousness instance has played no role whatsoever in any *thought* ever had about the consciousness instance.

Were this the case, there would be no cause to believe any argument made in support of the existence of the consciousness instance – not even by one's self – since the argument would have been made with indifference to the actual existence of the consciousness instance!

3.3.2 Possibility (II): active consciousness instances

The only other reasonable[3] possibility is that consciousness instances somehow *played an active role* in the aforementioned causal chain. In other words, that the ink molecules' equations of motion may not be fully understood without taking consciousness instances into account.

3.3.3 Evaluating possibilities (I) & (II)

We shall again reiterate: though possibility (I) appears much more reasonable given common wisdom of the brain (and indeed, of the universe), it is tragically flawed; it ultimately amounts to a *denial of the existence*

[2]Quasi-classical structures are those structures which are not *fundamental* to a classical model, and may not even be fully *accounted for* by a classical model, yet are *describable* using classical abstractions in the domain considered. An example is a modern CPU: though it is constructed of transistors ultimately accounted for only by quantum mechanics, the CPU's behavior in domains relevant to digital computing is describable by classical logic gates.

[3]Barring a cosmic conspiracy of unimaginable magnitude.

of the consciousness instance – on grounds of information *gleaned from the consciousness instance.* The ideas of a (3+1) dimensional universe, of the existence of particles, of physical structures abstracted as "brains" – are all revealed to us through the structure and relations within the consciousness instance. They thus cannot be used to justify the claim that the consciousness instance does not exist. Indeed, possibility (I) is *internally inconsistent*: it implies a conclusion which, were it true, would invalidate the assumptions which led to the conclusion in the first place.

Our hand is thus forced; we must accept possibility (II) as our only viable way forward.

3.4 Implications

The implications of this deceptively simple argument are staggering:

1. A complete and internally-consistent model of the universe must describe structures identifiable with consciousness instances, which further, must participate in the time evolution of structures we identify as "particles". In other words, **the consciousness instance is measurable – and therefore amenable to scientific examination.**

2. A-priori, one could expect the effects of the consciousness instance to be buried arbitrarily deep down the array of physical abstractions (e.g. in yet-to-be-discovered sub-sub-quantum physics). The above argument ties the "consciousness-matter" interaction to a common, specific, phenomenon: the vocalization of the thought "my body is associated with a consciousness entity". Thus **if a certain abstraction proves sufficiently powerful to model the brain's cognitive operations underlying the vocalization, it must *somehow* effectively capture the "consciousness-matter" interaction**.

There is a certain expectation in the physics community that the Hard Problem can forever be swept under the rug; that physics can describe the entirety of the universe without revealing any insight into the regularities governing the consciousness instance. This attitude has likely evolved, at least in part, due to the great difficulty of as much as envisioning a resolution for the Hard Problem. Nevertheless it has taken on a life of its own, leading to widespread implicit and even explicit denials of the Hard Problem – and thus to a near complete lack of research towards its resolution. Yet here we have a simple argument requiring of a consistent model of the universe not only to incorporate consciousness instances – but also to specify a measurable interaction between consciousness instances and "physical matter". If the attainment of a fundamental model of reality is truly our aim, we evidently have no choice but to confront the matter of the consciousness instance head-on.

Part II

Towards a Kinematical Theory

4 A Consciousness-Compatible Model of the Universe

"In this process, which constitutes the essence of scientific research, the human spirit neither hesitates nor fears to doubt the most self-evident facts of visual perception and to declare them to be illusions, but prefers to resort to the most extreme abstractions rather than exclude from the scientific description of nature one established fact, however insignificant it may seem."

— Max Born, *Einstein's Theory of Relativity*

Let us quickly recap our chain of reasoning thus far:

1. The consciousness instance exists.
2. Much regularity underlies the structure of each consciousness instance, strongly suggesting the existence of a universe external to the consciousness instance. Therefore, we posit, the "physical universe" exists.
3. "One's" consciousness instances bare a unique correlation to a specific structure in the "material universe": one's brain. Therefore we posit that there are structures in the "material universe" whose behavior is associated with consciousness instances.
4. If we accept (1-3), then the thought "the consciousness instance exists"[1] (or at the very least, an articulation of the thought) correlates with a physical process taking place in one's brain. This implies that the "material world" contains information originating with consciousness instances.
5. Therefore a model of the universe which accounts for the consciousness instance must allow for a *2-way interaction* between the consciousness instance and matter.

Evidently our search for a resolution of the Hard Problem has turned into a **search for a physical model** which would account for the known laws of physics – as well as convincingly model the consciousness instance.

4.1 New physics, or new interpretations?

We of course already possess a rather detailed model of the world around us – one which makes no reference whatsoever to the consciousness instance. Does our existing model simply break down in the face of the consciousness instance-"matter" interaction, or have we merely failed to properly identify the consciousness instances which our model already effectively accounts for?

4.1.1 The case for new physics

Though disagreement between theory and experiment has yet to be encountered under conditions thought to be relevant to cognition in the brain, it is conceivable that such disagreement would indeed be revealed on closer examination. In other words, that a detailed examination of the neural activity leading up to (a vocalization of) the thought "the consciousness instance exists" would reveal that neurons behave in manners not consistent with the known laws of physics.

No matter how seemingly unlikely, we cannot outright dismiss this possibility. We must remember, for instance, that were it not for *gravity's* peculiarly of being a purely attractive "force" (all particles on Earth attract a falling object in roughly the same direction), we might have failed to recognize *its* effects experimentally at this time; certainly, we would have been unable to detect said effects in experiments carried out in particle colliders. Should we conclude that existing physical theories are fundamentally incompatible with consciousness, a search for new physical effects in the brain would be utterly sensible.

What could such hypothetical new physics look like? In analogy with past advances in physics, one could imagine a "consciousness field" of sorts which is practically unmeasurable under conditions captured

[1] Not to mention the thought "a substructure of the consciousness instance currently associated with my brain is partially isomorphic to a rectangle".

by physics experiments carried thus far, but which is highly significant to the time evolution of particles in the brain. Thus our current physical model of the universe would be relegated to an *effective approximation* in the regime in which "consciousness field"-induced particle accelerations go to 0.

4.1.2 The case for new interpretations

While "unknown unknowns" could certainly catch the physics community by surprise, neither theoretical nor experimental considerations seem to hint at new physics in domains relevant to cognition in the brain. Though theoretical considerations lead us to *expect* the eventual failure of our existing models, we expect such failure to manifest only at very high energies – which seem entirely irrelevant to cognition in the brain. Experimentally, we have verified the predictions of our current models in domains which – on the surface of it – adequately capture the conditions relevant to cognition in the brain.

Thus we must carefully consider the possibility that the effects of the "consciousness-matter interaction" are effectively captured by our existing physical theories, and yet we have somehow failed to recognize this.

4.2 A minimal characterization of a consciousness-modeling structure

How is the hypothetical "consciousness field" described above different from the electromagnetic field? If the two aren't fundamentally different, then who's to say that consciousness instances *don't* correspond to that which is designated as "the electromagnetic field" in our theories?

Of course, incompatibilities (I) and (II) between the consciousness instance and classical models of the universe apply to the electromagnetic field as readily as to the neural model of the brain. Therefore the electromagnetic field – like our hypothetical consciousness field – helps us not at all.

If we are to develop a model of the universe which accounts for the consciousness instance we must finally confront our implicit guiding question: ***what would an acceptable consciousness-instance-modeling structure look like?*** Let us answer this question at the crudest level possible, identifying minimal constraints which must be satisfied by such a structure:

(I) **The consciousness structure must be affected by the effectively classical universe**
Since all that we know is the consciousness instance, and we merely conjecture that this consciousness instance mirrors in some fashion the "material world out there", this criterion is equivalent to the requirement that the "material universe" exists. Without this criterion, we can progress no further than *cogito ergo sum*.

(II) **The consciousness structure must affect the effectively classical universe**
As we have argued in (3), this point is required for a consistent model of the universe in which (I) applies. If the consciousness instance were purely passive with respect to the "physical" world, not an utterance nor a thought claiming to reflect on the nature (or the existence) of the consciousness instance would in actuality do so. Thus without this requirement, we would be forced to conclude that the consciousness instance does not exist: a cognitively unstable position.

(III) **The consciousness structure must carry fundamentally integrated classical information**
This requirement explicitly resolves incompatibility (I) between the consciousness instance and classical models of the universe. "Fundamentally integrated" is given by our earlier definition.

(IV) **The consciousness structure cannot be fully specified by classical information**
This requirement explicitly resolves incompatibility (II) between the consciousness instance and classical models of the universe. We assume that there are aspects of the consciousness instance which simply *cannot be written down*, but are necessary to fully characterize any particular consciousness instance (and therefore the "consciousness-matter interaction", and the time-evolution of our universe).

(V) **Implicit assumption: the consciousness structure is non-redundant**
This implicit requirement underlies all of our previous requirements. If the information (classical or otherwise) existing in a consciousness instance were fully determined by the consciousness-external universe, we would be wise to employ Occam's Razor so as to slice consciousness out of our theory altogether. Thus an account of the consciousness instance must be *required* for a determination of the time-evolution of the effectively-classical universe.

(VI) **Bonus points: the consciousness-classical interaction may non-trivially facilitate computation over classical information**
Unlike requirements (I-V), this requirement does not logically follow from our most basic notions of the universe; instead, it is a conjecture which seems reasonable given the circumstances in which we know consciousness to manifest.

We associate the consciousness instance with the state of a peculiar entity of the effectively-classical universe: the brain. The brain is of course peculiar not due to its mechanical properties, but due to its marvelous ability to facilitate computation. In fact, this peculiarity is so outstanding that some today believe that consciousness is somehow a side-effect of simply carrying out a complex computation in our universe.
Recognizing this peculiarity, we expect the interaction between the consciousness instance and the effectively classical universe to non-trivially facilitate computation over information which *is* fully specified classically (despite the condition specified by requirement (IV)).

It is difficult[a] to imagine what "non-trivial facilitation of computation" might mean in a world in which the extended Church-Turing thesis applies. More on this point, later.

[a] Though possible.

4.3 The Hard Problem revisited

In a universe containing structures which satisfy our consciousness constraints, the Hard Problem all but disappears. In such a universe, we would be able to specify regularities relating the effectively-classical universe to consciousness instances – in a manner consistent with our (normally implicit) requirements of a consciousness-instance-modeling structure.

Thus to finally lay the Hard Problem to rest, we must come up with a model of the universe which is consistent with our consciousness constraints, and experimentally verify that it indeed describes our own universe. Furthermore, we must verify that consciousness instances described by this model at the very least non-trivially participate in those causal chains taking place in one's brain which conclude with vocalizations of the thought "my brain is associated with consciousness instances".

5 Properties of a Consciousness-Compatible Universe

"In general, we look for a new law by the following process: first we guess it – no, don't laugh, that's really true; then we compute the consequences of the guess to see what would be implied if this law that we guessed is right; then we compare the result of the computation with observation, to see if it works."

— Richard Feynman, *recorded lecture*

5.1 Properties (II, IV) $\implies$ apparent (operational) indeterminism

As elegantly put by Hugh Everett[5], the question of whether the universe is *fundamentally* deterministic or indeterministic is not one that can be settled through experiment:

> *The question of determinism or indeterminism in nature is obviously forever undecidable in physics, since for any current deterministic [probabilistic] theory one could always postulate that a refinement of the theory would disclose a probabilistic [deterministic] substructure, and that the current deterministic [probabilistic] theory is to be explained in terms of the refined theory on the basis of the law of large numbers [ignorance of hidden variables].*

However irrespective of the *fundamental* nature of the universe, we shall see that a consciousness-compatible universe must *appear* to be indeterministic past a certain level of theoretic refinement. If the universe is fundamentally indeterministic this claim follows trivially; therefore let us suppose that the universe is fundamentally deterministic and demonstrate that this claim follows.

The intuition behind the claim is simple: if a component of the universe's state *cannot be written down*, then states which appear to be identical to one another (based on properties which *can be* written down), may not in actuality be so. Such seemingly-identical states would thus undergo divergent time evolutions – even in a deterministic universe.

Property (IV) tells us that any substructure S_j of our universe would (in general) consist of substructure specifiable by classical information, as well as of substructure which is not specifiable by classical information. Let us write:

$$S_j = (C_j, G_j)$$

where

- (C_j) denotes the classically-specifiable substructure of S_j
- (G_j) denotes the classically-unspecifiable substructure of S_j

Let us consider 2 systems, (I) and (II) given by the distinct yet related initial states ${S_i}^I = (C_i, {G_i}^I)$ and ${S_i}^{II} = (C_i, {G_i}^{II})$. The deterministic regularities of the universe would dictate that ${S_i}^I$ and ${S_i}^{II}$ evolve into the (generally distinct) final states ${S_f}^I = ({C_f}^I, {G_f}^I)$ and ${S_f}^{II} = ({C_f}^{II}, {G_f}^{II})$.

If we suppose that a scientific measurement can record only the classically-specifiable substructure of a given system[1], then even perfect measurements of systems (I) and (II) would reveal merely that:

- System (I) evolved from the state C_i to the state ${C_f}^I$
- System (II) evolved from the state C_i to the state ${C_f}^{II}$

[1] This is certainly true of every measurement recorded in a scientific journal.

By property (II), ${C_f}^I \neq {C_f}^{II}$ in general. Thus an examination of those measurements would suggest that a system initially given by the state C_i may evolve into either one of 2 distinct final states.

In other words, in a consciousness-compatible universe, *apparently* identical initial states may evolve into a number of distinct final states. Time evolution described thus is the very definition of (apparently) indeterministic time evolution.

5.2 Property (III) $\Longrightarrow$ departure from classical locality

According to property (III), the complex classical information embedded in a single consciousness instance must be *fundamentally integrated*. Let us recall the following condition on fundamental information integration specified in our earlier discussion (section 2.3.1):

> For information to be regarded as fundamentally integrated *with respect to a given model*, it should be impossible to state the regularities of the model in terms of subsets of the information.

Thus for the classical information embedded in a single consciousness instance to be regarded as fundamentally integrated with respect to our model *of the universe*, it should be impossible to specify the regularities *of the universe* in terms of subsets of said information.

This condition is far from innocent, and indeed marks a clear departure not only from our existing model of the brain – but possibly from *classical locality* itself.

♣ ♣ ♣

The regularities ("laws") of a local universe may be stated in terms of information associated with infinitesimally small regions of space-time (indeed, that is the very definition of locality). Consequently **in a local universe, it is only information associated with an infinitesimally small region of space-time which may be considered fundamentally integrated**. Consolidating this condition with the requirement that the complex information of a single consciousness instance be fundamentally integrated presents us with considerable – seemingly insurmountable – difficulties.

Firstly, classical relativistic theories associate only a *small, finite* number of parameters with each infinitesimal region of space-time in any given reference frame (e.g. a particle's mass, its electric charge, the magnitude and direction of its velocity, the components of the electromagnetic four-potential, etc.). While (in most models) each such parameter p is given by a *real* number $r_p \in \mathbb{R}$ – which in principle may encode arbitrarily many bits of information – in practice, even utilizing advanced experimental techniques, we find that:

- for all intents and purposes, we can recover only a modest amount of bits from each such parameter, due to the difficulty of precisely measuring its value
- for all intents and purposes, we can encode only a modest amount of bits into each such parameter, due to the difficulty of precisely altering its value

As such, it boggles the mind to imagine that the brain somehow manages to reliably encode and decode the (comparably rich) information embedded in a single consciousness instance into and out of the available local parameters.

Furthermore, even if we assume for a moment that the complex information making up a single consciousness instance could somehow be encoded into and decoded out of an infinitesimal region of space-time, it is difficult to imagine that such an infinitesimal consciousness instance could meaningfully affect the state of the brain so as to facilitate complex cognitive operations (such as a vocalization of the thought "I am associated with consciousness"). An infinitesimal consciousness instance would essentially behave as a particle (or point in a field), which could not be reasonably expected to meaningfully affect the behavior of the comparably vast brain[2].

♣ ♣ ♣

Thus it seems that to seriously entertain the notion of the fundamental integration of the consciousness instance's complex substructure, we must be prepared to abandon locality itself.

[2]If pressed, one could perhaps imagine a holographic brain of sorts, in which countless, nearly-identical local consciousness instances interact with one another and with the rest of the brain so as to bring-about coherent cognitive operations which effectively incorporate the information common to all such instances. Though qualitatively comparable schemes in which global

5.3 For the second time: new physics or new interpretations?

A century ago, requiring of a consciousness-compatible model of the universe to exhibit the aforementioned characteristics would have seemed to doom our efforts altogether. The universe described by (special & general) relativity and by Maxwell's equations – our best models of the universe at the time – is explicitly deterministic and local, and thus can support neither classically-indescribable states nor complex fundamentally integrated classical information. Indeed, explicit determinism and locality (and by extension, separability of complex information) were seen as defining features of any physical ontology, to an extent more fundamental than the particulars of any given theory.

And yet (as is well known) these characteristics – *apparent indeterminism*, and *fundamental integration of complex information* (along with *non-locality*) are not only apparently compatible with our universe after all – they are the very hallmarks of our best-yet model of the universe: quantum mechanics. These characteristics are *precisely* that which sets quantum mechanics apart from classical physics, and which has been so difficult to fully make sense of over more than a century of study.

By now our intension is surely clear. We propose that the effects of the consciousness instance have indeed been measured in the laboratory, and have been dubbed *quantum mechanics*.

information is encoded into and subsequently recovered from countless local instances do in fact exist (e.g. a standard visual hologram), in the context of the operations of the brain such a scheme seems too fantastic to seriously entertain.

We will note, however, that the prospect may seem somewhat less fantastic in the context of theories under which it is somehow "easy" to encode and decode complex information into and out of infinitesimal regions of space-time – as is the case with some entirely-local interpretations of QM (such as variants of the Many-Worlds interpretation). We will not explicitly pursue this direction of inquiry any further.

Part III

Consciousness and Quantum Mechanics

6 A Consciousness-Compatible Quantum Ontology: Motivations and Hopes

"Physics does not consist only of atomic research, science does not consist only of physics, and life does not consist only of science. The aim of atomic research is to fit our empirical knowledge concerning it into our other thinking. If it cannot be fitted into [this other thinking], then it fails in its whole aim and one does not know what purpose it really serves."

— Erwin Schrödinger

6.1 Motivations for a Consciousness-QM correspondence

With the advent of the quantum revolution came the realization that many of our intuitive assumptions about the nature of reality do not, after all, coincide with the behavior of our universe. Since the incompatibility between those intuitive assumptions and our notions of consciousness had long been recognized (if only implicitly), speculations of non-trivial connections between consciousness and QM soon surfaced.

Such speculations have tended to revolve around the suggestion that the (apparent) *collapse of the wavefunction* exhibited by quantum systems may be triggered by an interaction between a "material" system and a "conscious" observer. Notoriously, such proposals have not only led nowhere – they have also, at times, attracted the attention of mystics and crackpots not engaged in true scientific exploration.

Many in the physics community have reacted to this predicament with decidedly disproportionate push-back, declaring all efforts to find a connection between QM and consciousness to be inherently misguided and illogical. It is often sarcastically asked: *quantum mechanics is weird, and consciousness is weird, therefore they must be related?*

We must now set the record straight: deep motivation for a consciousness-QM correspondence may be found in but a cursory survey of our assumptions. As we argued in chapter 3, an *internally-consistent* model of reality must describe a 2-way interaction between consciousness structures and effectively-classical physical systems. It is thus eminently clear that a model which *fails* to account for consciousness structures cannot be a fundamental one. QM is of course regarded as a model capturing our universe's *fundamental* underlying structure, rather than as a merely useful calculation tool capturing important characteristics of particular physical systems (unlike, for instance, the Ising model). We should therefore expect QM to be intimately related to the matter of consciousness.

Furthermore, even if QM is one day usurped by a more refined theory of nature, existing evidence gives us good reason to expect QM to *effectively* capture the operations of the brain *relevant to cognition*. Since the consciousness instance must be compatible with laws of physics merely *complete enough* to effectively capture said operations, it is reasonable to expect consciousness to be compatible with QM – *even if* QM turns out to capture but an effective description of our universe's structure. Indeed, if QM were shown to effectively capture the brain's cognitive operations (as it is expected to) and yet to be unequivocally incompatible with consciousness – *our entire picture of the universe* would arguably prove cognitively unstable.

To such elementary motivations we may of course add our more refined considerations. We were led to an explicit examination of a consciousness-QM correspondence by recognizing in QM 2 properties which we deemed foundational to any consciousness-compatible model of reality: effective indeterminism and non-locality. If QM itself *does not* effectively describe consciousness structures, this striking relation must be entirely coincidental, and further, QM must merely approximate a more refined model of our universe's structure which *does* describe consciousness structures – and which *independently* manifests those 2 quantum-mechanical properties.

To be sure, the search for a consciousness-QM correspondence is further motivated by the underlying confusion ("weirdness") surrounding quantum theory – for it is hoped that such a correspondence would

shed light not only on consciousness, but also on QM itself. However we must again stress that we would be justified in searching for a consciousness-QM correspondence even if quantum theory were perfectly and uncontroversially understood.

6.2 The *non-relativistic* foundation for a consciousness-compatible quantum ontology

Though a more fundamental relativistic description exists, we shall, for the time being, consider QM only in the non-relativistic limit. To reiterate once again, since non-relativistic QM is expected to *effectively* capture the operations of the brain relevant to cognition, it should likewise *effectively* capture the effects of consciousness. Furthermore, it is noteworthy that those features of non-relativistic QM which are most remarkable (and which we shall shortly discuss) have made it unscathed into relativistic QM.

The well-known *effective* postulates of non-relativistic QM are:

1. A physical system is perfectly characterized by $|\Psi\rangle$: a ray in a Hilbert space, normalized to unity.
2. The system evolves in time according to the Schrodinger equation:
   $\hat{H}\Psi = i\hbar\frac{\partial\Psi}{\partial t}$
   where $\hat{H}$ is the system's Hamiltonian.
3. a. Each physical observable is associated with a Hermitian operator on the Hilbert space.
   b. When making a *measurement* of $|\Psi\rangle$ using a Hermitian operator $\hat{T}$, such that $\hat{T}$ has eigenvalues $\{t_i\}$ associated with eigenvectors $\{|t_i\rangle\}$, one will get eigenvalue t_i with a probability given by $|\langle t_i|\Psi\rangle|^2$.
   c. Following such a measurement, the system originally described by $|\Psi\rangle$ will *collapse* to a state given by the eigenstate associated with the measured eigenvalue.

We dub the above *effective postulates* due to the notoriety of the 3^{rd} postulate, which – while useful in practice – is in principle inconsistent with the 2^{nd} postulate. If the measurement apparatus is a structure described by QM, then it never "measures" any given system. Instead, the "measurement apparatus" and the "measured system" form a composite quantum system which should, by postulate (2), unitarily evolve in time in accordance with the Schrodinger equation.

6.3 Quantum peculiarities

The relations between the more peculiar qualities of QM seem to hint at a deeper and simpler underlying description yet to be revealed. That an explicitly consciousness-compatible account of QM could offer such a description hardly seems out of the question.

Let us then draw attention to some of the well-known peculiarities of (effective) QM. They shall soon play a role in our construction of a consciousness-compatible quantum ontology.

6.3.1 The measurement problem

The so-called measurement problem refers to the ambiguity introduced by the 3^{rd} effective postulate of QM. We will defer to John Bell for an elucidation [6, *Quantum mechanics for cosmologists*, p. 124]:

> *The problem is this: [effective] quantum mechanics is fundamentally about 'observations'. It necessarily divides the world into two parts, a part which is observed and a part which does the observing. The results depend in detail on just how this division is made, but no definite prescription for it is given.*

When Bell says *"The results depend in detail on just how this division is made"* he means that experiments would, in principle, provide different results depending on where *exactly* in the process collapse occurs. To see this, consider for instance that the time evolution of a wavefunction before its collapse is unitary and therefore reversible, while the time evolution of the wavefunction "during" collapse is irreversible.

Of course, the results depend *in detail* on how the division is made only *in principle*. In practice, it is sufficient to push the so-called "Heisenberg cut" dividing the observer from the observed far enough from the observed such that no detectable differences in results would manifest if the cut were moved any further. This "good enough" prescription places the cut firmly in the microscopic realm in virtually all practical settings.

It has become increasingly clear in recent decades that much of that which has been traditionally attributed to the 3^{rd} effective postulate naturally arises out of postulates (1-2). In particular, the (emergent) phenomenon of *quantum decoherence* shows that even *barring a collapse of the wavefunction*, the composite system formed by a microscopic subsystem entangled with a macroscopic environment in a certain sense behaves indistinguishably from a composite system in which the microscopic subsystem has collapsed.

6.3.2 Quantum indeterminism

According to the effective quantum formalism, identical systems may not, in general, behave identically over time; in particular, the behavior of systems deemed identical to one another may diverge upon their "measurement". This divergence occurs because effective QM can only predict a measurement's result with certainty when the measured system is given by an eigenstate of the measurement-associated operator. In all other cases, effective QM can merely predict the *statistical distribution* of measurement results (when carried over an infinite ensemble of such supposedly-identical systems).

The ultimate source of QM's indeterminism has been a focal point of much (perhaps too much, if we consider Everett's remarks) of the discussion surrounding QM's foundations. Have we correctly identified the fundamentally probabilistic regularities relating substructures of our universe to one another? Or have we failed to account for all aspects of our universe's structure, and thus failed to recognize the fundamentally deterministic regularities relating substructures of our universe to one another?

6.3.3 The Born rule

The Born rule is the name given to the particular probabilistic prescription of measurement results predicted by effective QM: the probability that a measurement of a physical system given by $|\Psi\rangle$ would yield eigenvalue λ_i is given by $|\langle\lambda_i|\Psi\rangle|^2$. It has long been wondered whether this rule can be somehow *derived* from more fundamental principles.

6.3.4 EPR: indeterminism-driven quantum non-locality

As first identified by Einstein, Podolsky, and Rosen, effective QM's account of indeterminism is intimately related to a peculiar non-locality inherent to the quantum formalism.

Under effective QM, a measurement of a physical system is posited to instantaneously bring-about the "collapse" of its entangled counterparts – no matter the space-time separation between them. Since the "collapse" of a system affects its time evolution, interaction with one (apparently-localized) physical system is in essence posited to affect a substructure of the universe relevant to the ("immediate") time evolution of *arbitrarily distant* (apparently-localized) physical systems – violating the spirit of the explicitly-local relativity.

Confusingly, this non-local coupling between entangled systems manifests in practice only in the form of correlations between measurement results – and cannot be used to facilitate the transmission of information between distant systems. Furthermore, such effects **naively** seem to be *at the very least compatible* with behavior arising out of pre-arranged correlations, rather than out of non-local interactions.

As demonstrated by EPR's arguments[7], should quantum mechanical experiments become understood without an appeal to such non-local interactions, they would necessarily also become understood without an appeal to the flavor of indeterminism pointed to by effective QM.

♣ ♣ ♣

EPR went on to conclude that effective QM is likely to merely approximate a more fundamental model explicitly manifesting locality – and hence also determinism[1].

Though Einstein is widely regarded to have lost touch with physics by rejecting the completeness of quantum theory, upon reflection it rather boggles the mind that Einstein's position with respect to QM *was in the minority*! EPR's rejection of QM's completeness did not originate with blind belief in a transcendent principle of aesthetics governing the regularities of the universe (as is suggested by Einstein's all-too-quotable "God does not play dice"). Instead, it originated with EPR's recognition of the enormity of the non-locality claim, as well as of its seemingly miraculous – hence "spooky" – (apparent-) unverifiability.

Though a so-called "hidden variables" account of QM's predictions was not immediately available, why did most physicists of the previous century assume that one could not be discovered?[2] Is the existence of undiscovered degrees of freedom truly less reasonable than a world of cats superposed between life and death, and of "spooky actions at a distance"?

6.3.5 Bell's theorem: explicit quantum non-locality

Arguably the most profound discovery in the history of physics was presented by John S. Bell in an almost painfully simple 5-page paper titled *"On the EPR Paradox"*[8]. In this paper, Bell showed that *no theory* of **local** hidden variables could reproduce the statistical predictions of quantum mechanics – no matter how complex a local model we are willing to admit. In essence, Bell showed that the *regularity* relating spin measurement results of maximally-entangled (spatially-separated) spin-$\frac{1}{2}$ particles *could not* arise out of local regularities operating on information associated with the immediate space-time surroundings of each particle in isolation.

We have provided a brief review of Bell's (ultimately simple) theorem[3] in Appendix A. Let us also comment on the spirit of Bell's argument.

♣ ♣ ♣

All physical models are in a sense calculation tools used to predict measurable properties of physical systems. *Fundamental* physical models harbor an additional goal: to *maximally* describe the underlying structure of said physical systems.

As agreement between theory and experiment improves, it becomes easy to conflate experimental confirmation of a theory's calculable predictions with validation of its account of the fundamental nature of reality. However if we take a step back, it should become plain that any model we examine, no matter how well in agreement with experiment, may capture merely an *effective* description of our universe's underlying structure.

Keeping this point in mind, let us draw our attention to the physical model we call QM. This model has certain striking characteristics about it; wave-particle duality, mutually incompatible reality of non-commuting observables, indeterminate collapsing measurements, "spooky actions at a distance", etc. Those are characteristics of *QM the calculation tool*.

It is natural to wonder what conclusions, if any, we may draw from the agreement between experiment and *QM the calculation tool* with regards to the *fundamental* nature of reality. In other words, suppose we *assume* that the quantum-mechanical model of a physical system captures merely an effective description of the physical system's (arbitrarily rich) underlying structure; may we confidently attribute *any* qualities to said underlying structure, other than to say it is effectively modeled by QM under suitable conditions?

[1] At least in the context of effective QM's entanglement. EPR's arguments left the door open for indeterminism of local interactions, e.g. during the formation of entangled particles.

[2] Of course, various hidden-variables (so-called) "impossibility theorems" were published over the years, most infamously von Neumann's. However in the words of John Bell: *"the von Neumann proof, if you actually come to grips with it, falls apart in your hands! There is nothing to it. It's not just flawed, it's silly... When you translate [his assumptions] into terms of physical disposition, they're nonsense. You may quote me on that: The proof of von Neumann is not merely false, but foolish!"* (Interview in Omni, May, 1988, p. 88). Furthermore, considering the fact that Bohm's work achieved the supposedly "impossible", it seems that physicists of the previous century were all too hasty in their acceptance of QM's completeness.

[3] In its pedagogically superior form originally described by Greenberger-Horne-Zeilinger.

That, fundamentally, was the question answered by Bell in his seminal paper. The answer offered by Bell's theorem, then, shines light on that which may be regarded as the *essence* of the quantum-mechanical model of reality: **given a system of entangled particles, the substructure of physical reality corresponding to a measured value of one particle cannot be independent from the substructure of reality affected by a measurement of its entangled counterparts** – no matter the space-time separation between the particles.

6.3.6 No-communication theorem

It may appear as though Bell's theorem opens the door for gross violations of classical causality by allowing faster-than-light communication between distant regions of space-time. However the no-communication-theorem shows that in the context of effective QM, the (local) interaction between an "observer" and a particle cannot be used to transmit information to an "observer" with (local) access to another particle – even in the presence of entanglement between the 2 particles.

The regularities modeled by entanglement relations in effective QM therefore occupy an interesting space: they cannot arise out of local regularities operating on information associated with infinitesimal regions of space-time, and yet they do not admit transmission of information, as would be the case for naively-constructed non-local regularities.

6.3.7 No-cloning theorem

The no-cloning theorem states that in the context of effective QM, it is impossible to create an identical copy of an arbitrary unknown quantum state.

6.3.8 Quantum teleportation

Though it is impossible to clone a quantum state, it is possible to transmit a quantum state from one system to another by transmitting only classical information. The protocol used to achieve this involves the scrambling of the original quantum state, and makes use of an entangled EPR pair whose constituent particles have been pre-shared between the origin system and the destination system.

6.3.9 Holovo's theorem and superdense coding

A qubit is a structure much more complex than a bit. The state of a single qubit is in general specified by 2 real numbers (capturing its orientation on a Bloch sphere); the state of n qubits is in general specified by $2^n - 1$ complex numbers. One may thus reasonably expect n qubits to carry much more information than n bits. However Holovo's theorem shows that in the context of effective QM, no more than n classical bits may be extracted from n qubits.

Despite Holovo's bound on the extractable information content of a single qubit, superdense coding schemes allow for the transmission of 2 classical bits through the transmission of only a single qubit. Such schemes make use of an entangled EPR pair whose constituent particles have been pre-shared between the origin system and the destination system (similarly to quantum teleportation).

Let us keep those motivations and peculiarities in mind as we head towards the formulation of a coherent consciousness-accommodating physical ontology.

7 Towards a Coherent Ontology

"What is proved by impossibility proofs is lack of imagination."

— John S. Bell, *On the Impossible Pilot Wave*

Our aim in this chapter is to take the first steps towards our ultimate goal: the formulation of a coherent physical ontology accounting for the known laws of physics (in the form of effective QM), as well as for consciousness. We will strive to make headway by simultaneously pulling on both ends of the emerging QM-consciousness correspondence, searching for structures which illuminate each of the correspondence's poles, as well as bridge the gap between them.

7.1 Side (I) of the correspondence: the classical indescribability and fundamental integration of consciousness instances

Other than identifying 2 of its qualities which he hold to be classically-incompatible, we have thus far called little attention to the nature of the consciousness structure. Instead, we have relied on rough intuitive notions evoked by not much more than our crude working definition. If we are to model the consciousness structure, we must first sharpen our conception it – and of its classically-incompatible qualities in particular.

7.1.0 Foreword: the analysis of consciousness

When it comes to the analysis of consciousness, we have but a single tool at our immediate disposal: introspective observation. It may seem odd – perhaps even out of place – to employ introspective analysis towards the formulation of a model of the universe of which we demand objectivity and rigor. After all, science (and physics in particular) typically deals only with those examinations of reality which are ultimately captured by the positions of lab instruments' dials. By giving weight to introspective observations we seem to lose not only the clarity and simplicity afforded by the position of a lab instrument's dial, but also its inherent undeniable objectivity.

To such sentiment we must respond with a question: who among us has ever directly inspected a lab instrument's dial? With some reflection it should become clear that one never directly inspects any element of reality other than one's current consciousness instance. The positions of instrumentation dials are always *inferred* from the contents of consciousness instances – which are assumed, through the mediation of our senses, to (approximately) mirror the substructure of reality referred to as "the positions of the instrumentation dials". The introspective analysis of consciousness thus constitutes the ultimate, inescapable cornerstone of all physical investigation.

That is not to say that objectivity is forever lost. When assessing the position of a lab instrument's dial, all conscious entities ultimately arrive at the same conclusion (e.g. "the dial is pointing at the number 4"), though each draws on the structure of "his/her" consciousness instance. In other words, **conscious entities are able to agree upon the substructure of "their" consciousness instances corresponding to the positions of instrumentation dials**. It therefore seems at the very least hopeful that subtler aspects of the consciousness structure would likewise prove amenable to consensus upon sufficient reflection.

Another important quality of the position of a lab instrument's dial is that it coherently relates to our wider cognitive framework. A lab instrument consists of atoms characterized by mass, momentum, electric charge, etc. – much like the objects whose properties it measures. Thus the *lab instrument itself* coherently fits within the model capturing the system *measured by* the lab instrument. As we move towards the formulation of a coherent, falsifiable, consciousness-accommodating model of the universe, **we should likewise expect those subtler aspects of the consciousness structure to (eventually) coherently fit into (and hence be corroborated by) our model of the wider universe.**

7.1.1 The classical-indescribability of consciousness instances

We typically assume that the classically-indescribable facet of consciousness is — in a sense – redundant. We take it for granted that for any given individual, the correlation between (classically-describable) brain states and (classically-indescribable) qualia remains consistent over time, concluding that a quale could not capture any causally significant information *not already captured* by a classical description of its associated brain state.

For instance, suppose that Alice is gazing at electromagnetic radiation approximately 700 nm in wavelength – colloquially known as "red light". Some of the structure of Alice's brain state is then captured by the *classically-describable* label *"Alice is looking at red light"*. **How does this label relate to the (partially-classically-indescribable) consciousness instances associated with Alice's brain?**

Alice maintains that whenever her brain activity is well-described by the label *"Alice is looking at red light"*, consciousness instances associated with her brain contain *a particular range* of (classically-indescribable) qualia. She therefore designates said qualia range as "**the** red-associated qualia" (for her own brain). Alice may even call said qualia "red", using a single label ("red") to characterize: (1) the (conjectured) electromagnetic radiation which triggered her visual system, (2) the state of her (conjectured) "material" brain, and (3) those (classically-indescribable) qualia supposedly common to all consciousness instances corresponding to "red-detecting" brain states. In doing so, Alice implicitly asserts that sensory input, classically-describable brain states, and classically-indescribable qualia always (or at the very least, typically) exist in similar configurations.

Since Alice believes that "**the** red-associated qualia" are associated with her brain activity if and only if her brain is well-characterized by the classically-describable label *"Alice is looking at red light"*, she concludes that the nature of those qualia cannot capture any casually significant information *not already captured* by the label describing her brain state. In other words, Alice maintains that qualia are passive substructures of our universe somehow existing "in parallel" to classically-describable substructures of our universe, and that the particular quale manifesting in any given consciousness instance could not factor into the time evolution of our universe.

♣ ♣ ♣

In the same breath, we often wonder whether all individuals perceive identical (or at leasts similar) color-associated qualia when perceiving identical sensory input. For instance, Alice may wonder whether the quale which she calls "red" is called "blue" by Bob, and indeed, whether her color-associated qualia overlap with Bob's at all.

♣ ♣ ♣

Let's carefully examine the assumptions implied in the above analysis.

First and foremost, in asking whether 2 qualia are identical to one another, we imply that **qualia – while classically-indescribable – may be *compared* against one another and judged to be similar or dissimilar** (for instance, while hiking in nature, we judge the color-associated qualia corresponding to green grass to be similar to the color-associated qualia corresponding to the green leaves of a tree).

We take this assumption to be self-evident. In a sense, it is a postulate inherent to our most basic notion of consciousness.

♣ ♣ ♣

Having accepted that qualia may be compared against one another, we further surmise that for any given individual, brain states specified by identical classically-describable parameters will be accompanied by consciousness instances consisting of identical classically-indescribable qualia (e.g. the "same" classically-indescribable qualia will forever correspond to those objects we label as "red").

This second assumption smells less like a postulate and more like a *recognition of a regularity.* The regularity being: *"whenever I gaze at light given by a certain classically-describable distribution of wavelengths (colloquially called 'red light'), I experience identical classically-indescribable color-associated qualia".*

What have we observed which led us to identify this regularity? We tell ourselves: *"I remember the quale which I associated with the label 'red' yesterday, and it is identical to the quale which I associate with the label 'red' today".*

Let us take a step back and carefully consider the above statement. Yesterday's red-associated quale was a component of *yesterday's consciousness instance*. How, then, can we remember yesterday's red-associated quale? Do we truly have access to yesterday's consciousness instance? If we do not, then what precisely do we mean when we say that we "remember" yesterday's red-associated quale?

Upon reflection, it should become clear that on recalling a memory, a representation of yesterday's events (which has been recorded in one's brain state) suddenly becomes reenacted within one's *current* consciousness instance. Though we naively believe that we compare yesterday's red-associated quale with today's, in actuality we simply compare the quale encoded in the past-associated substructure *of the current consciousness instance* with the quale encoded in the present-associated substructure of the current consciousness instance. Indeed, it may not be a bad idea to regard one's experience of the past as a "memory sense", on the token of regarding one's eyesight as a "(present) photon sense", or one's hearing as a "(present) air-vibration sense". Thus the close similarity between the qualia encoding one's *memory of the color red* and one's *sensory perception of the color red* is neither more fundamental nor more interesting than the close similarity between the qualia encoding the color red in one's *left-eye-associated field of vision* and in one's *right-eye-associated field of vision*.

♣ ♣ ♣

We can thus appreciate that **we only ever compare qualia against one another *across a single consciousness instance*.** Despite strong intuitions suggesting the contrary, we do not *truly* have access to yesterday's consciousness instance – nor to last minute's.

Though this conclusion does not *immediately* imply that the qualia associated with classically-identical[1] brain states are not *also* identical to one another, we can appreciate that we have neither evidence nor need for requiring our ontology to associate classically-identical brain states with identical qualia. Indeed, applying the principle of Occam's razor, it would be wiser to require that in general, this condition does not hold true.

Recalling our assertion that consciousness instances must participate in the time-evolution of the effectively-classical universe, we can thus appreciate that **the *particular* classically-indescribable quale manifesting in association with a given brain state may well be causally significant – as it need not redundantly correlate to the classically-describable brain state**.

♣ ♣ ♣

Our intuition that yesterday's "red"-associated quale is identical to today's "red"-associated quale is so strong, that the notion that the 2 *may not* in fact be identical is difficult to accept – even after accepting each step in the above chain of reasoning.

It may be rendered a little less jarring by recognizing a certain likeness between this new conception of qualia and our experience of *music*. In general, we recognize a tune not by the absolute pitches of the notes played, but by the *relations between those notes*. If a tune is played $\frac{2}{7}$'s of an octave higher than usual, listeners still easily recognize it as the same tune. Though every one of the notes played is quite different from the original rendition's, *the relations between the notes* remain identical.

This *likeness* between our new conception of qualia and our perception of music is not an *equivalence* because in addition to remembering the *relations* between pitches of notes, we also have a certain capacity to remember the *absolute pitches* of notes we hear[2]. Nevertheless recognizing this likeness may leave our proposed "relativity of qualia" a little more familiar, and a little easier to digest.

7.1.2 The fundamental integration of consciousness instances

The above analysis reveals an intricate relationship between the classical-indescribability of a consciousness instance and its fundamental integration: *comparisons* between classically-indescribable qualia are only possible when the qualia make up a single (fundamentally integrated) consciousness instance. In other words: **classical information is typically extracted out of classically-indescribable qualia by comparing**

[1] Judged to be identical based on classically-describable depictions.

[2] Some more so than others, but we can all at the very least tell that a very high note is not a very low note.

the qualia making up a single (fundamentally integrated) consciousness instance against one another.

It is tempting to consider the stronger claim that *all* classical information found in a single consciousness instance is encoded in relations between classically-indescribable qualia. For the time being, we shall leave this claim as a conjecture to be revisited in the future.

7.2 Side (II) of the correspondence: indeterminism and non-locality in effective QM

7.2.1 Bell's theorem

Let us consider 2 maximally-entangled spin-$\frac{1}{2}$ particles in the Bell state:

$$\left|\Phi^{+}\right\rangle = \frac{\left|0\right\rangle_{I}\left|0\right\rangle_{II} \quad + \quad \left|1\right\rangle_{I}\left|1\right\rangle_{II}}{\sqrt{2}}$$

Without loss of generality, let us draw our attention to results of spin measurements of particle (I).

When viewed in isolation from particle (II), effective QM cannot predict the result of a spin measurements of particle (I) with probability better than chance – no matter the basis in which the spin measurement is carried out. Thus the result of particle (I)'s spin measurement, when viewed in isolation from particle (II), is *maximally uncertain* according to effective QM.

However when conditioned on the result of particle (II)'s spin measurement, effective QM predicts a drastically different (probabilistic) distribution of measurement results: particle (I) will be found to be in the $\left|1\right\rangle_{I_\theta}$ state with probability given by $\cos^2\frac{\theta}{2}$, where θ is the angle between the axis along which particle (I)'s spin is measured, and the axis along which particle (II)'s spin was found to be in the $\left|1\right\rangle_{II_\theta}$ state.

♣ ♣ ♣

As previously discussed, Bell's theorem conclusively proves that no (reasonable) *local* model of the universe could account for such behavior. Thus to account for observed experimental results, we must accept that particles (I) and (II) in the above Bell state are *somehow* non-locally coupled, such that **the process of measurement of particle (II)'s spin affects the state of the universe relevant to the determination of particle (I)'s spin measurement result**[3] (and vice-versa).

Furthermore, the non-local coupling between entangled particles is neither random nor uniform across experiments, as different experiments are associated with **quantitatively distinct** non-local couplings. The *relative, probabilistic* characterization of a particular experiment's non-local coupling is given by the structure of the wavefunction ascribed to the experiment by effective QM. In other words, **to explain the results of experiments described by effective QM, one must posit that the state of the universe *somehow* encodes structure which is associated not with a point in (effective) spacetime, but rather with certain (effective) particles, and which must – at the very least – somehow encode the *relations* between the particles' measured properties**.

7.2.2 No-communication theorem

Though Bell's theorem forces upon us the assumption that the process of measurement carried-out on particle (II) changes a component of the state of the universe relevant to the time evolution of particle (I) – the no-communication theorem suggests that said change cannot be used to transmit information from the environment of particle (II) to the environment of particle (I).

Having introduced the notion of classically-indescribable state, we should now be more specific. The no-communication theorem suggests that said change cannot be used to transmit *classically-describable* information from the environment of particle (II) to the environment of particle (I).

7.2.3 Classical-indescribability and indeterminism

We have previously pointed to the (partial) *classical-indescribability* of consciousness instances as the likely *source of the effective indeterminism* characterizing physical systems described by QM. We have thus suggested that the underlying regularities governing the time evolution of those physical systems may well be fundamentally *deterministic*.

[3] At least from the "reference frame" of the "observer" which conducted the measurement on particle (II).

How does the notion of a deterministically evolving system sit with the notion of a classically-indescribable system?

Suppose we have before us a physical system modeled under QM as a spin-$\frac{1}{2}$ particle in the $\left|+_{\hat{Z}}\right\rangle$ state. If the regularities of the universe are fundamentally deterministic, then the results of spin measurement of the particle along *any* given axis are predetermined. We could thus imagine a function laying out precisely said results[4]. Wouldn't such a function constitute a classically-describable account of all of that is measurable about the particle's state?

Indeed, such a function *would* provide a classical account of all measurable aspects of the system – **were the experiment's *final state* completely classically-describable**. Why should we care about an unobserved, supposed classically-indescribable component of our experiment's final state? In the real universe, the so-called "final state" of our experiment may itself serve as the initial state of a subsequent experiment. This relation of course carries-on recursively, with each experiment's final state constituting a future experiment's initial state. Thus to fully capture the state of our initial system, our imagined function would have to account not only for the result of our initial experiment – but also for the results of all *dependent* experiments.

Keeping Bell's theorem and the no-communication theorem in mind, our emerging ontology seems to suggest that **in the case of an *entangled* multi-partite system, the process of measurement carried-out on one local component affects the classically-indescribable state of the universe relevant to the time evolution of the *other* (entangled) components**.

In our review of EPR's arguments (6.3.4), we drew attention to the intricate relationship between the indeterminism and the non-locality characterizing effective QM. In (7.1.2) we drew attention to the subtle relationship between the consciousness structure's classical indescribability and its fundamental integration. It is therefore perhaps not too surprising to find that our emerging ontology attributes QM' indeterminism to the consciousness structure's classical-indescribability, and its non-locality to the consciousness structure's fundamental integration.

[4]Potentially a highly complex, chaotic function.

7.3 An emerging correspondence

7.3.1 2 doubly-flawed models

Let us now explore 2 (highly) simplified models, each capturing a different aspect of our emerging ontology. Though both models are fatally flawed, the manners in which they fail shall prove instructive.

(I): classically-indescribable but not fundamentally integrated

We can imagine a universe containing classically-indescribable structures which can be assigned a classically-describable relation to other classically-indescribable structures. Said classically-indescribable structures would be associated with (structures otherwise corresponding to) particles, such that the time evolution of a particle may depend on the particular classically-indescribable structure it is associated with.

On the consciousness side of the correspondence, the classically-indescribable structures would of course correspond to qualia. On the QM side of the correspondence, the association between particles and classically-indescribable qualia would account for the effectively-indeterministic time evolution of particles.

For instance, suppose that particle A is associated with the quale corresponding to the label "red" in your current consciousness instance, while particle B is associated with the quale corresponding to the label "blue" in your current consciousness instance. Since the particles are associated with different qualia, we would expect their time evolutions to eventually diverge – even if their states are otherwise identical.

♣ ♣ ♣

This scheme seemingly provides us with an interpretation for quantum entanglement: entangled particles would correspond to particles associated with (distinct) classically-indescribable qualia – which have a known (classically-describable) relation between them (for instance, 2 particles may be known to be associated with identical qualia – though the particular nature of said qualia would be classically-indescribable and hence unknown). Thus a measurement of one of the particles would teach us about a corresponding measurement of the other.

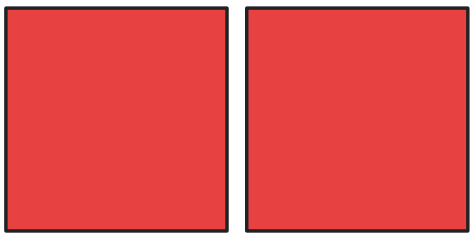
Having recognized that our descriptions of consciousness instances (e.g. "I am seeing the color red") essentially boil down to descriptions of *relations* between classically-indescribable qualia ("the quale associated with visual sensation in this consciousness instance is identical to the quale associated with the word 'red' in this consciousness instance"), descriptions of consciousness instances would thus correspond to the substructure of wavefunctions describing entanglement relations.

For instance, consider the Bell state $|\Phi^+\rangle = \frac{|0\rangle_I|0\rangle_{II} + |1\rangle_I|1\rangle_{II}}{\sqrt{2}}$. We would interpret this state as corresponding to a pair of particles associated with classically-indescribable (and thus effectively unknown) qualia – which are nonetheless *known to be identical to one another*. Such a state would presumably also correspond to an (imaginary) consciousness micro-instance consisting of 2 identical color-associated qualia (figure 7.1).

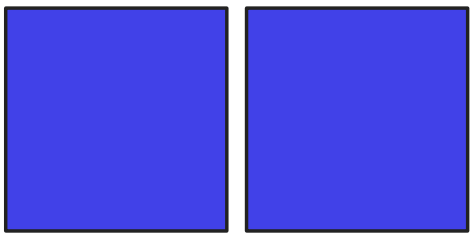
Likewise, we would interpret the Bell state $|\Psi^+\rangle = \frac{|0\rangle_I|1\rangle_{II} + |1\rangle_I|0\rangle_{II}}{\sqrt{2}}$ as corresponding to a pair of particles associated with qualia which – while individually unknown – are known to affect the behavior of a particle in "opposite" manners. Such a state would presumably also correspond to an (imaginary) consciousness micro-instance consisting of 2 "opposite" color-associated qualia (figure 7.2).

♣ ♣ ♣

Though possibly appealing at first glance, this scheme ultimately breaks down as it fails to capture the fundamental information integration inherent to effective QM as well as to consciousness. On the QM side, Bell's theorem of course teaches us that the behavior of an entangled particle pair could never arise out of pre-determined correlations, as is the case with the aforementioned scheme. On the consciousness side, this scheme provides us with no means of identifying the qualia (and hence the information) making up a single consciousness instance; in a world of independent classically-indescribable qualia "floating about", how would we tell which qualia make up a single consciousness instance, and hence what information is encoded in the relations of a single consciousness instance?



(a) A symbolic representation of a consciousness micro-instance consisting of 2 identical color-associated qualia, where each is identical to the color-associated quale labeled as "red" in your current consciousness instance.



(b) A symbolic representation of a consciousness micro-instance consisting of 2 identical color-associated qualia, where each is identical to the color-associated quale labeled as "blue" in your current consciousness instance.

Figure 7.1: 2 symbolic representations of (imaginary/simplified) consciousness micro-instances which may be well-described by the Bell state $|\Phi^+\rangle = \frac{|0\rangle_I|0\rangle_{II} + |1\rangle_I|1\rangle_{II}}{\sqrt{2}}$. Each consciousness micro-instance consists of 2 classically-indescribable qualia which are identical to one another. The 2 consciousness micro-instances, while structurally similar to one another, are in fact distinct from one another – accounting for the divergent time evolutions of systems described by the Bell state $|\Phi^+\rangle$.

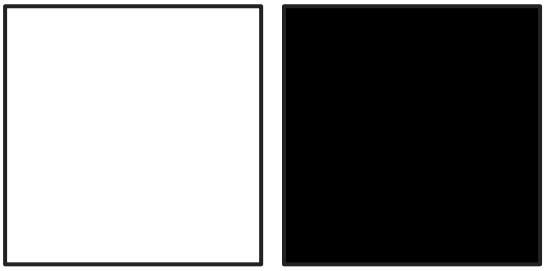


Figure 7.2: A symbolic representation of (an imaginary/simplified) consciousness micro-instance which may be well-described by the Bell state $|\Psi^+\rangle = \frac{|0\rangle_I|1\rangle_{II} + |1\rangle_I|0\rangle_{II}}{\sqrt{2}}$. One of the qualia is identical to the quale labeled as "white" in your current consciousness instance; the other is identical to the quale labeled as "black" in your current consciousness instance.

(II): fundamentally integrated but not classically indescribable

In the context of physical theory, an example for the fundamental integration of complex information can be found in an ideal rigid body evolving in time in a Newtonian universe.

An ideal rigid body is a 3-dimensional structure described by a time-invariant spatial mass distribution (given in a reference frame attached to the body), as well as by time-variant linear and angular positions capturing the body's position and orientation in a given inertial reference frame.

Since an ideal rigid body cannot be broken down, it cannot meaningfully be said to be made of *independent* constituent components. Furthermore, to calculate the time evolution of an ideal rigid body, the full description of the body must be taken into account, along with a consideration of all forces acting – and not acting[5] – on the body. As such, we may reasonably consider the information describing an ideal rigid body to be fundamentally integrated – for the regularities of the universe in which it operates cannot be stated in reference to subsets of said information.

♣ ♣ ♣

In the context of our discussions, we could imagine consciousness instances as rigid-body-like structures which may be associated with particles. Entanglement would thus arise out of a rigid-body-like connection between entangled particles.

♣ ♣ ♣

Though such a scheme would indeed capture the fundamental information integration of a single consciousness instance, it would fail to capture its classically-indescribability. Furthermore, on the QM side, such a scheme would fail to account for the effective indeterminism exhibited by quantum systems, as well as open the door for violations of the no-communication theorem (and therefore for causality-breaking superluminal communication of classical information).

7.3.2 An emerging ontology

While far from fully formulated, a coherent ontology begins to emerge in a synthesis of sorts between the 2 aforementioned models.

We may roughly think of consciousness instances as kinds of semi-rigid bodies, consisting of classically-indescribable "points" (qualia) with a "rigid" classically-describable relation between them. A classically-indescribable quale may be associated with a particle – rendering its time evolution effectively indeterministic. Certain kinds of interactions with the particle would invariably "rotate" and "bend" the consciousness instance, affecting the positions of *all* of its constituent "points" (qualia), as well as the relation between them; interactions with one particle may therefore affect its entangled counterparts, as well as the entanglement relation itself.

Though the *particular* quale associated with a given particle is classically-indescribable and hence effectively unknowable, the *classically-describable relations* between different qualia may be known and even predictable – and may well manifest in experiments in the form of *correlations* between the results of localized measurements. Since consciousness "rotations" and "bending" would only affect the *classically-indescribable* state relevant to a distant particle's time evolution, they could not be used to facilitate the transmission of classical information.

7.3.3 A brief summary

This point bears repeating. Let us take this opportunity to briefly summarize our conjectures along with our emerging ontology.

We propose that the consciousness instance be regarded as a 1st-class[6] primitive of physical reality – whose relation to other 1st-class primitives of physical reality is subject to a certain set of regularities. We have identified 2 central differences between the structure of consciousness-instance-modeling-objects and the familiar objects considered in classical models:

[5] Counterfactual definiteness?

[6] At the very least initially.

1. The consciousness instance harbors complex internal structure; the regularities relating a given consciousness instance to other physical primitives are stated in terms of this complex structure *in its entirety* – deeming the structure fundamentally-integrated.

2. The consciousness instance's structure is only partially classically-describable; nevertheless the classically-indescribable substructure of the consciousness instance participates in the (*potentially deterministic(!)*) regularities relating consciousness instances to other physical primitives.

We further propose that the above structure and behavior is *effectively* exhibited by physical systems modeled under quantum mechanics:

1. Operators act on a given wavefunction as a whole. Per Bell's theorem, this behavior is fundamental to our physical reality.

2. Physical systems whose classically-describable structures are maximally identical (i.e. physical systems described by identical wavefunctions) may undergo divergent time evolutions under certain circumstances (i.e. wavefunction collapse).

These conjectures manifest in the context of our emerging ontology, as we can demonstrate by once again considering (imaginary/simplified) consciousness micro-instances associated[7] with elements of physical reality which are effectively modeled as particles.

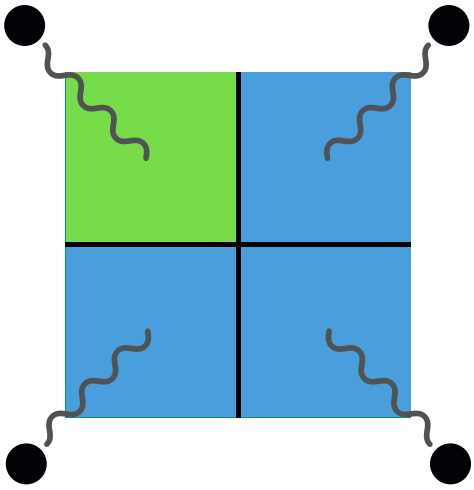
Take some particle system X consisting of the particles $\{p_1, p_2, p_3, p_4\}$.

- Under our criteria, we may suppose that whenever particle system X is associated with the consciousness micro-instance corresponding to (7.3a), a certain interaction (Int) with p_1 will drive system X towards an association with the consciousness micro-instance corresponding to (7.3b).

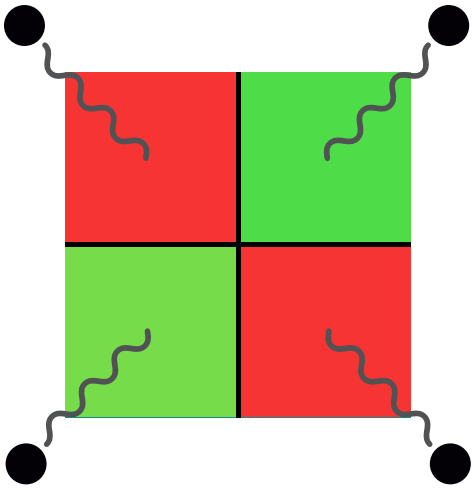
- We may suppose further that whenever particle system X is associated with the consciousness micro-instance corresponding to (7.3c), the same interaction (Int) with p_1 will drive system X towards an association with the consciousness micro-instance corresponding to (7.3d).

Note that the consciousness micro-instances corresponding to (7.3a) and to (7.3c) differ from one another **only in their classically-indescribable substructure**, while the consciousness instances corresponding to (7.3b) and to (7.3d) differ from one another in their **classically-describable** substructure as well.

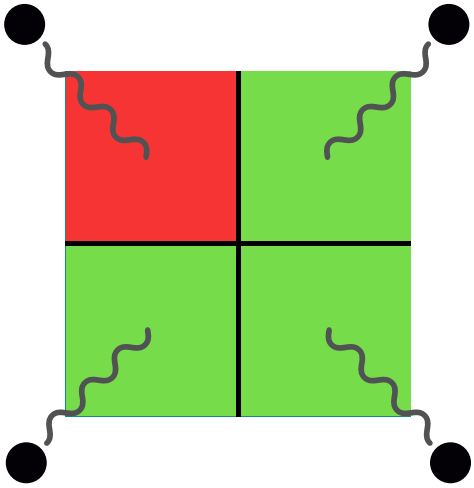
[7] By "associated" we mean: to the extent to which some substructure of physical reality is captured by a certain particle system, a consciousness instance is associated with the particle system if *it effectively affects the behavior of* said particle system, and is *effectively affected by* said particle system. For example, "one's" current consciousness instance is associated with the particle system making up one's brain.



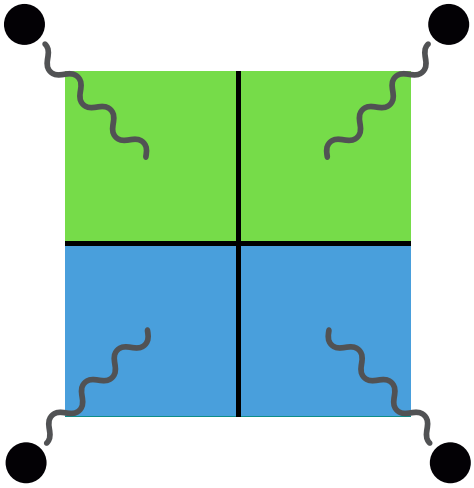
(a) A representation of a consciousness micro-instance whose classically-describable substructure is equivalent to the classically-describable substructure of the one given by (7.3c).



(b) A representation of a consciousness micro-instance whose classically-describable substructure is distinct from the classically-describable substructures of all other consciousness micro-instances referenced here.



(c) A representation of a consciousness micro-instance whose classically-describable substructure is equivalent to the classically-describable substructure of the one given by (7.3a).



(d) A representation of a consciousness micro-instance whose classically-describable substructure is distinct from the classically-describable substructures of all other consciousness micro-instances referenced here.

Figure 7.3: Examples of (imaginary/simplified) consciousness micro-instances, illustrating our emerging ontology. We use color to represent the classically-indescribable substructure of the consciousness instance; a green square, for instance, represents not some generic classically-indescribable parallel to the concept of "green", but the classically-indescribable quale corresponding to the color green in your *current* consciousness instance. As such, only the classically-describable substructure of the consciousness instances may be written down – the substructure capturing the *relations* between a consciousness instance's classically-indescribable qualia.

7.4 Quantum peculiarities in light of our emerging ontology

Our incomplete emerging ontology can already be seen to suggest particular interpretations to some of QM's puzzling peculiarities.

7.4.1 Quantum indeterminism

The state of a physical system is (in general) not fully isomorphic to classical information – and is thus (to an extent) unknowable. The time evolution of an arbitrary physical system is consequently rendered (to an extent) unpredictable.

7.4.2 Bell's theorem

The state of the universe relevant to the time evolution of a given system is indeed affected by an interaction with its distant entangled counterparts. In other words, the regularities of our universe are explicitly non-local.

7.4.3 The no-communication theorem

The state of the universe which is subject to non-local regularities is classically-indescribable (on all ends of the interaction). Thus non-local interactions cannot be used to facilitate the transmission of classical information.

7.4.4 Quantum teleportation

Let us define the notion of *state-class equivalency*. Under our emerging ontology, a quantum state in fact represents an entire *state-class* of (partially classically-indescribable) physical states. Thus under our ontology, 2 physical systems which are described by the same quantum state may not be truly identical. Nevertheless we may say that the 2 systems are *state-class equivalent* to one another – i.e., their physical states are well-described by the same quantum mechanical state-class.

Let us now review the quantum teleportation protocol, and then consider it in light of our emerging ontology:

First, an EPR pair is prepared and shared between Alice and Bob. Let us consider the Bell state given by $|\Phi^+\rangle_{AB} = \frac{|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B}{\sqrt{2}}$. Suppose that Alice wishes to transmit to Bob the quantum state given by $|\psi\rangle_C = \alpha|0\rangle_C + \beta|1\rangle_C$.

Our initial quantum state is thus given by

$$|\Phi^+\rangle_{AB} \otimes |\psi\rangle_C = \frac{|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B}{\sqrt{2}} \otimes (\alpha|0\rangle_C + \beta|1\rangle_C)$$

which can be shown to be equivalent to

$$\begin{array}{l} \frac{1}{2}\left[|\Phi^+\rangle_{AC} \otimes (\alpha|0\rangle_B + \beta|1\rangle_B) + |\Phi^-\rangle_{AC} \otimes (\alpha|0\rangle_B - \beta|1\rangle_B)\right. \\ \left.+|\Psi^+\rangle_{AC} \otimes (\beta|0\rangle_B + \alpha|1\rangle_B) + |\Psi^-\rangle_{AC} \otimes (\beta|0\rangle_B - \alpha|1\rangle_B)\right] \end{array}.$$

Alice then "measures" the A-C qubits in the Bell basis, "collapsing" the composite ABC system into one of the four quantum states (each with equal probability):

1. $|\Phi^+\rangle_{AC} \otimes (\alpha|0\rangle_B + \beta|1\rangle_B)$
2. $|\Phi^-\rangle_{AC} \otimes (\alpha|0\rangle_B - \beta|1\rangle_B)$
3. $|\Psi^+\rangle_{AC} \otimes (\beta|0\rangle_B + \alpha|1\rangle_B)$
4. $|\Psi^-\rangle_{AC} \otimes (\beta|0\rangle_B - \alpha|1\rangle_B)$

This measurement requires a local interaction between the physical systems associated with the A and C qubits.

The final state of the composite system is always given by one of the states (1-4). Furthermore, from the result of her measurement of the A-C-subsystem, Alice may deduce which of those 4 possible states indeed corresponds to the final state of the composite system.

Alice may thus transmit to Bob 2 classical bits encoding the description of the final composite system, allowing Bob to perform local operations on his qubit[8] so as to bring it into a state identical to the $|\psi\rangle_C$ state.

From the point of view of our ontology, we may think of the measurement process as a collision of sorts between the origin system C and the "semi-rigid, partially classically-indescribable body" associated with the pre-shared EPR pair AB. This collision confers upon the remote component of the EPR pair – the destination system B – a new classically-indescribable state related to its own initial state as well as to the initial state of the origin system C.

We should not be surprised to find that the post-collision *classically-describable relation* between the AC qubits can be inferred by Alice, who has access to both. Since the relation between the AC qubits is related to their mutual relation to the B qubit, that too can be inferred by Alice.

Alice can then share her findings with Bob, revealing to him the state-class to which his qubit's state belongs (post-collision). Bob can then use said information to transform the state of the B qubit into a state well-described by the state-class $|\psi\rangle_B = \alpha|0\rangle_B + \beta|1\rangle_B$; in other words, into a state which is state-class equivalent to the initial state of the origin system C [9].

7.4.5 Superdense coding

Once again, let us review the quantum protocol before considering it in light of our emerging ontology:

An entangled EPR pair is shared between Alice and Bob. Like before, let us consider the entangled state given by $|\Phi^+\rangle_{AB} = \frac{|0\rangle_A|0\rangle_B \ + \ |1\rangle_A|1\rangle_B}{\sqrt{2}}$.

Alice may perform on her A qubit any one of the operations $\{\mathbb{I}, \sigma_3, \sigma_1, \sigma_3\sigma_1\}$. The application of such an operation shifts the composite system to a state given by $\{|\Phi^+\rangle_{AB}, |\Phi^-\rangle_{AB}, |\Psi^+\rangle_{AB}\ |\Psi^-\rangle_{AB}\}$, respectively.

Alice may then send her A qubit to Bob, who may measure the AB system in the Bell basis, and deduce which of the 4 operations was in fact carried out by Alice.

Let us now consider this protocol under our emerging ontology. The *particular* classically-indescribable states (qualia) associated with the 2 components of our pre-shared EPR pair are unknown; nevertheless there is a known "rigid" relation between the components of the pair. By manipulating one of the particles making up the EPR pair, we may manipulate its rigid relation to its counterpart. Different manipulations result in different final rigid relations between the particles.

Following our manipulation, both particles are associated with new classically-indescribable states which are, again, unknown (and classically unknowable). However the *relation* between those 2 states is known, and depends on the particular manipulation operation carried out. The *relation* between the 2 classically-indescribable components *is* classically describable – and is thus, in principle, knowable. It is therefore not surprising to find that a procedure exists which explicitly uncovers the *relation* between said components – thereby revealing which of the 4 manipulation procedures was originally carried out.

[8] Operations $\{\mathbb{I}, \sigma_3, \sigma_1, \sigma_3\sigma_1\}$ for cases {1, 2, 3, 4}, respectively.

[9] Note that quantum teleportation schemes between an origin system C and a destination system B must merely cause the system B to assume a final state which is *state-class equivalent* (rather than identical) to the initial state of system C.

7.5 Comparisons against The Everett / Relative-State / "Many Worlds" interpretation of QM

Our emerging ontology seems to describe a universe characterized by certain qualitative attributes. Let us compare this universe against the universes described by the MWI of QM.

The MWI is rooted in the conjecture that the 3rd effective postulate of QM is entirely superfluous. That is, that all observed behaviors of quantum systems emerge out of the unitary evolution of the wavefunction as prescribed by the Schrodinger equation – including all aspects of the apparent collapse of the wavefunction.

7.5.1 Decoherence

The MWI interpretation is typically considered in conjunction with the phenomenon of the **decoherence of the wavefunction** – which is well-established to indeed emerge out of the unitary evolution of the wavefunction as prescribed by the Schrodinger equation.

Decoherence describes the process through which the (composite) wavefunction describing a microscopic system in interaction with a macroscopic "environment" tends to evolve towards state of superposition between distinct "branches" – which experience vanishing (mutual) interference effects, and which correspond to quasi-classical "worlds". Furthermore, each such quasi-classical "world" is weighted in proportion to the probability assigned to the emergence of said world by the Born rule[10].

7.5.2 The Many-Worlds ontology

Since the MWI takes the position that the collapse of the wavefunction simply never takes place, all "branches" of the wavefunction are regarded as equally real, and forever continue to evolve in accordance with the Schrodinger equation.

Though disturbing at first glance, much about the MWI is ultimately appealing. The universe described by postulates (I) and (II) of effective QM indeed seems to contain within it structure isomorphic to the world we see around us. If it *also* seems to contain structures isomorphic to worlds which we do not see around us – but which we could have conceivably seen (had certain quantum measurements yielded different results) – then so what? After all, the suppression of interference effects between the branches even accounts for the fact that we see no direct evidence of such supposedly parallel worlds.

7.5.3 The origin of indeterminism

The MWI clearly paints a *qualitatively* indeterminate picture of ("subjective") reality. If "observers" correspond to certain quasi-classical substructures of the universe's wavefunction ("brains" manifesting in "branches" of the wavefunction"), and if branches associated with *all* possible results of a given measurement indeed manifest, then following a "measurement", our universe shall contain "observers" which have witnessed each possible measurement result. After a series of "measurements", our universe shall contain "observers" which have witnessed each possible *record* of measurement results. Since *all possible* measurement result records indeed manifest, then clearly no single regularity could account for the relation between the past/future substructures of *each*.

Let us restate the above in the language of consciousness. **The MWI effectively claims that consciousness instances are associated with substructures of the wave-function which we would identify as (time slices of) quasi-classical brains.** Following a "measurement", consciousness instances

[10] Earlier, we noted that in most practical settings, the Heisenberg cut dividing the quantum mechanical "observed" from the "observer" may be placed firmly in the microscopic realm without coming into conflict with the results of (current-day) experiments. The phenomenon of quantum decoherence shows that the converse is also true: the Heisenberg cut may be placed arbitrarily close the opposite edge of the von-Neuman chain coupling a system's state to one's experiences. In other words, observed phenomena naively appear to be *consistent* with the position that the universe's wavefunction undergoes "collapse" only when it is to manifest in the information content of a consciousness instance (and that only a single consciousness instance exists in the universe at any given time: "one's own"). We can perhaps appreciate how the topic of the psycho-physical correspondence has grown so entangled with the world of QM, and with the notion of wavefunction collapse in particular.

encoding the "observations" of *each possible measurement result* exist in the universe. After many such measurements, the universe shall contain consciousness instances corresponding to brains which have witnessed *each possible record* of measurement results. Therefore no single regularity could account for the relation between the past/future substructures of *each* of our universe's consciousness instances.

♣ ♣ ♣

The MWI does not as easily account for the *quantitative* nature of the indeterminism exhibited by quantum systems. To appreciate this, let us once again draw our attention to consciousness instances corresponding to "brains" which have amassed records of many quantum measurement results (to reiterate, per the MWI interpretation, said brains are represented by nothing more than quasi-classical substructures of the universe's wavefunction). For instance, let us consider brain structures recording the results of measurements of many qubits, each in the state $\frac{|0\rangle + \sqrt{2}|1\rangle}{\sqrt{3}}$.

Those brains shall identify all sorts of regularities in their respective measurement records: one brain will record a string of only 0's; another, a string of only 1's; another, an alternating string of 0-1-0-1-..., etc. Some of those brains will identify *statistical* regularities in their records: roughly half of the records of some brains will consist of 0's, with the other half consisting of 1's, in a manner compatible with measurements of $(\frac{1}{2} : \frac{1}{2})$ random variables. Indeed, some of those brains will identify statistical regularities in line with those predicted by the Born rule: results compatible with measurements of $(\frac{1}{3} : \frac{2}{3})$ random variables.

Let us not fool ourselves: if we make the claim that consciousness instances are associated with all quasi-classical "brainy" substructures of the universe's wavefunction – as the Everett interpretation implicitly asserts – then consciousness instances which encode the "identification" of each such *regularity* shall exist in our universe. What is so special then about Born-rule-compatible branches of the wavefunction?

It is true that we can identify non-trivial relations between the "contents" of Born-rule-compatible branches (the measurement records encoded in the branches) and their associated weights. For instance, the amplitude-squared sum of the of the weights of Born-rule compatible branches approaches 1 as we increase the number of qubit measurements considered[11]. However in attributing any significance to the weights assigned to different branches of the wavefunction, are we not already, in some fashion, accepting the Born rule?

If such relations fully illuminate the matter of the origin of the Born rule, than it is not at all well-established *how*.

7.5.4 Comparison against our emerging ontology

The most striking difference between the Many Worlds ontology and our own emerging ontology of course lies with their respective qualitative accounts of the indeterminism of experience.

Under the MWI, the root cause of the indeterminism of experience lies with the multitude of "experiences" which emerge out of each effective "measurement"; given a physical system well-modeled by quantum mechanics (a "quantum system") which is "measured" by a conscious entity, consciousness instances corresponding to "observations" of *each* possible measurement results manifest in our universe.

Under our emerging ontology, the root cause of the indeterminism of experience lies with our inability to account in our calculations for the classically-indescribable state of our universe; it may be too early to assert with certainty, but this seems to suggest that given a "quantum system" which is "measured" by a conscious entity, the consciousness instances which manifest in our universe following a measurement correspond to but a *single* measurement result.

♣ ♣ ♣

This of course suggests that our emerging ontology is ultimately in tension with the Schrodinger equation itself after all, for unlike the Schrodinger equation, our emerging ontology does not treat all (effective) branches of the wavefunction equally. Nevertheless our ontology is not only consistent with the predictions of effective QM to a good approximation – it also appears to maintain one of the Schrodinger's equation most profound and cherished characteristics: its unitarity.

[11] Though we can be more precise in this statement, the central point is hopefully clear.

Though *any degree* of disagreement with the predictions of the Schrodinger equation may gravely trouble some, we must appreciate that in seeking to *exactly* apply the Schrodinger equation to *the entire universe*, we extrapolate its validity far beyond its verified domain of applicability. Should we stake our entire picture of the universe on an equation which we have verified only in limited domains, which we cannot consolidate with our theory of gravity, and which we cannot as much as clearly interpret?

In assessing the Schrodinger equation, we must also address the elephant in the Everettian room. The entire Everettian scheme hangs on a conjectured correspondence between consciousness instances and "brain-resembling" substructures of the universe's wavefunction. Indeed, Everett himself clearly understood this, and explicitly referenced the psycho-physical correspondence no fewer than 3 times in his original dissertation. However the structure of the wavefunction does not satisfy all of our consciousness constraints; in particular, it does not support the notion of classically-indescribable state[12]. It is perhaps noteworthy that the consciousness constraint not satisfied by the structure of the wavefunction is the one associated with indeterminism – arguably the MWI's weakest suit.

[12]This critique would apply to virtually all other existing interpretations of QM.

8 Discussion

"All theories are insights, which are neither true nor false but, rather, clear in certain domains, and unclear when extended beyond these domains."

— David Bohm, *Wholeness and the Implicate Order*

8.1 Theoretical investigation to follow

Our journey is plainly far from its point of conclusion. A sensible next step would be the development of a mathematics of deterministically-evolving classically-indescribable structures, and the explicit uncovering of its relations to probability theory, as well as to other branches of mathematics. Such a mathematical framework would allow for the formalization of some of the more intuitive arguments laid out in this paper; the argument laid out in (7.2.3), for instance, suggests a certain conservation relation which such a mathematics may point to. Furthermore, it is not unreasonable to hope that the formalization of the assumptions of classical indescribability would pave the way for a satisfying derivation of the Born rule, perhaps through a certain relation to Gleason's theorem.

Armed with a mathematics of classically-indescribable structures, we may then proceed towards the formulation of a coherent consciousness-accommodating ontology accounting for the behavior of quantum mechanical systems described under the simplified assumptions of quantum information science. Given a physical system well-modeled by (spacetime-agnostic) qubits and quantum gates, such an ontology ought to explicitly lay out the (fundamentally integrated) classical information content of each consciousness instance associated with the system, as well as fully account for its partial classical indescribability (and thus for the emergence of the Born rule). As this ontology evolves, investigations of "consciousness mechanics" may be made through the emerging subfields of quantum information science.

Following the development of such an ontology, we may begin reformulating it to accommodate the more accurate assumptions of (non-simplified, spacetime-dependent) QM, and eventually, of QFT.

8.2 Verification, falsification, and further experimental investigation

8.2.1 Qualitative qualities of quantum mechanics

There is a sense in which our preliminary model "predicts" some of the most striking qualitative characteristics of quantum mechanics itself. In chapter (5), without alluding to any particular physical experiment (other than the "experiment" that is ordinary life), we were able to claim:

- That the universe must be effectively (i.e. operationally) non-deterministic past a certain level of theoretic refinement.
- That the universe must support complex, fundamentally integrated classical information. As we briefly mentioned, this requirement may be equivalent to a requirement of non-locality when considered alongside other reasonable requirements; even if it is not, it marks a clear departure from the standard classical picture.
- That the above 2 properties of the universe are likely to somehow support structures which "non-trivially" facilitate computation; in other words, that the Extended Church-Turing Thesis is likely to be incorrect when classically-indescribable, fundamentally integrated structures are involved.

In principle, these claims could have been made decades ago, if not hundreds of years ago. Indeed one can discern hints of the arguments we've covered in the writings of Descartes and of Leibniz (among others). Had

these claims been explicitly made prior to the discovery of quantum mechanics, *the discovery of quantum mechanics itself* would have surely served as *preliminary* verification of our ideas.

In practice, however, these claims are *so very striking* that despite mounting experimental evidence, even *quantum mechanics itself* was repeatedly rejected prior to the discovery of Bell's Theorem (and rightly so). One could hardly expect a scientifically-minded individual who stumbled upon the notions we have presented here *prior to the discovery of quantum mechanics* to have entertained them long enough so as to seriously develop them.

8.2.2 Experimental verification, a "smoking gun", and falsification

One of our central claims is that one cannot simultaneously believe that:

1. consciousness is real.
2. The brain's operations leading up to the articulation "consciousness is real" can be understood without accounting for consciousness.

This claim implies that to completely understand consciousness, one "merely" has to completely understand the brain's operations leading up to an articulation of the sentence "consciousness is real". In other words, the Easy Problem and the Hard Problem must be inexorably related.

It seems that to verify our preliminary ideas, we must find the classical information content of our consciousness instances to manifest in entanglement relations of complex quantum states arising in the brain.

Since the ability to analyze the brain in sufficient detail alludes us – and will likely continue to allude us for decades (if not centuries) to come – it will be useful to search for our emerging model's smoking gun: evidence that the human brain *somehow* facilitates complex quantum entanglement. That we naively expect the brain to be extremely inhospitable to non-trivial entangled states only serves to further strengthen this prediction's suitability as an initial test for our emerging ontology.

It is not our purpose to meaningfully speculate with regards to particular mechanisms which might support complex entangled states in the brain. A number of candidate mechanisms have been proposed over the years, including [9, 10]. The Posner molecule[11] seems to make for a particularly promising candidate.

Alternatively, to falsify our claims, it must be shown that the brain's cognitive operations leading up to an articulation of the utterance "consciousness exists" may be understood in terms of interactions between effectively classical subsystems.

8.3 The mirror test

In principle, our ideas are tantamount to the introduction of a third (and potentially final) layer to the "mirror test".

In the first and most basic layer of the "mirror test", one realizes that the 3-dimensional figure reflected in the image of a mirror is somehow intimately related to "one's" consciousness entity; its eyes record that which one's consciousness entity sees, its ears record that which one's consciousness entity hears, and its limbs move as "commanded" by one's consciousness entity. One then declares to one's self: *this structure reflected in the mirror is* ***me****.*

With the realization that the state of one's physical body corresponds to the state of one's consciousness entity comes a question: what is the substructure of one's body which directly relates to one's consciousness? It initially seemed as though consciousness somehow correlates to the body as a whole, rather than to any well-localized region within the body. After all, one's consciousness is manifestly correlated with the state of one's eyes, one's ears, one's limbs, etc.

Through *careful observation and experimentation* we eventually arrived at the second layer of the "mirror test". We learned that the contents of one's consciousness are correlated not with the state of one's eyes, ears, limbs, or kidneys – but only with the state of one's *brain*. Any correlation between consciousness and other parts of the body is today understood to be the result of the strong (nerve-facilitated) *coupling* between the state of the brain and the states of other parts of the body.

Modern-day neuroscience further describes correlations between substructures of the brain and substructures of consciousness instances. For instance, we can identify areas of the brain which correlate with the visual information content of consciousness, and areas of the brain which correlate with auditory information content of consciousness. Once again, the question arises: what is the substructure of one's brain which directly relates to one's consciousness? This question is of course nothing but a reformulation of the Hard Problem.

Our discussion essentially leads us to propose the addition of a third (and potentially final) layer to the "mirror test": the identification of consciousness instances with entangled quantum states manifesting in the brain.

Mirroring the transition from the first to the second layer of the mirror test, we predict that the the correlations between effectively-classical substructures of the brain and consciousness and will come to be understood in terms of strong *couplings* between the effectively classical and quantum realms.

8.4 Consciousness and cognition

Though not entirely worked out, our ideas seem to suggest that complex consciousness is a hallmark of *quantum computation*. If that is indeed the case, our preliminary model strikingly also satisfies our "bonus" constraint for a consciousness-modeling structure[1].

8.5 How many consciousness instances are associated with a single brain?

It is typically implicitly assumed that a single brain is associated with but a single consciousness instance during any given moment. This assumption is in no way self-evident, and must be subject to experimental investigation. For instance, it is possible that peripheral consciousness instances are involved in the (so-called) subconscious processes which our "primary (talking) consciousness" interacts with.

8.6 On the matter of "free will"

Free will has long been a topic of philosophical discussion. Though the ideas we have presented here do not introduce fundamentally new concepts into the relevant philosophical arena, it might interest some to review the matter through our new lens.

8.6.1 The free will regularity

Discussions of free will often emphasize the "free" in "free will" – which is typically taken to mean "unconstrained". The question then immediately follows: what exactly is this "will" unconstrained *from*? Our model of reality is ultimately a codification of certain universal regularities ("the laws of physics"). Is "free will" not constrained by these regularities? If that is the case, then in what manner does it fit into our model of the universe?

We shall tap into the discussion at a point a bit further upstream, with a question: *what drives one to claim to possess free will?* Surely, the claim is not the result of an extensive empirical study of the regularities of nature and of the operations of one's brain.

Ironically, the claim of having "free will" in actuality arises out of the *observation of a regularity*: whenever we think to ourselves "I wish to do X", X is carried out! Any account of free will, then, must simply explain this regularity, and nothing more.

[1] That is, barring a jarring discovery that BQP == BPP.

8.6.2 Active consciousness

A matter we have thus far alluded to only in passing is that consciousness seems intimately tied to the aforementioned "free will regularity". Indeed, a more precise statement of the regularity is that whenever "one's"*current consciousness instance* contains the thought "I wish to do X", X is subsequently carried out.

Since consciousness has traditionally been thought of as a passive structure (when not outright considered to be an illusion), this regularity has been viewed as a problem to be contended with. It has been suggested, for instance, that though we intuitively think the "free will regularity" to be the result of causation, it is in fact merely the result of (indirect) correlation; that the consciousness manifestation of "I wish to do X" arises out of the same mechanism that causes X, but does not actively participate in the causal chain resulting in X being carried out.

However when consciousness is regarded as an *active* structure – a structure which must be accounted for in a calculation of the universe's time evolution – the problem with the "free will regularity" all but disappears. The thought "I wish to do X", can reasonably be said to indeed cause X.

It is important to remember that consciousness constitutes merely one of many elements taking part in the causal chain leading up to the carrying-out of X, and that quasi-classical structures also participate in this chain. Therefore we should not be surprised to find correlations, or even causal relations between such quasi-classical structures (e.g. classically describable neural activity) and the carrying out of X. For instance, we shouldn't be surprised to find that (consciousness-blind) MRI scans can predict which of 2 buttons a subject will press with probability better than chance.

8.6.3 Agency vs. freedom

It is interesting to note that there is nothing "free" about the account of "free will" our model is pointing to; consciousness is no less constrained by the regularities of nature than the motion of the heavenly bodies, or the propagation of electromagnetic waves. Furthermore, our model conceivably robs the universe of the fundamental indeterminism (wishfully?) thought by some to harbor a fleeting possibility of a truly unconstrained will. Nevertheless, in an ironic twist, it seems that man's intuitive sense of *agency* could hardly have been more deeply affirmed than by our model.

8.7 The evolutionary origins of consciousness

If our consciousness model is correct, it is natural to wonder how conscious entities such as ourselves have come to inhabit the world. That *we have* is particularly remarkable given the fragility of complex quantum entanglement with respect to decohering effects; even more so considering the warm, wet environments that are biological systems.

Of course, no matter how unlikely the natural evolution of a conscious entity is, we should not be surprised to find *ourselves* (or our ancestors and relatives) to be conscious entities – based on standard (if slightly more refined) anthropic arguments: if we *weren't* conscious entities, there would be "nobody home" to wonder about such matters.

Nevertheless it is interesting to wonder about the likelihood of the natural evolution of conscious entities, in particular in relation to non-conscious life, and non-conscious intelligent life. It is also interesting to wonder whether consciousness is *necessary* for the practical realization of human-level intelligence, or if our own species' evolution of intelligence has, in a sense, been skewed by anthropic effects.

If we find that the natural evolution of consciousness is as unlikely as it naively appears to be, and further, that consciousness is an effective prerequisite for the practical realization of human-level intelligence through the processes of evolution, consciousness could conceivably factor into a "great filter" resolution of the Fermi paradox.

8.8 Premature questions and speculations

8.8.1 Awareness of consciousness itself

The author of this paper knows better than most that no definitive statement with regards to the nature of consciousness is made effortlessly. Our working definition for the consciousness instance as well as our minimal characterization of a consciousness-compatible universe were arrived at through much reflection over the course of many years. Nevertheless, I imagine that most readers will come to agree with the observations presented here upon sufficient reflection. Furthermore, while the incongruity between the classical universe and consciousness is difficult to state precisely, many can easily see that is indeed *somehow* there.

Us humans seem to have significant (if somewhat confused) cognitive access not only to the "external world" captured by consciousness – but also to consciousness itself. Why should this be?

On the one hand, an explicit account of this matter is not strictly required. We can imagine, as consciousness-deniers often do, that evolution has given rise to a cognitively-useful (and therefore evolutionarily advantageous) model of consciousness, which can be represented *in* consciousness. Unlike consciousness-deniers, imagining that this is indeed the case leaves us with an internally-consistent model of reality. We even have good reason to expect evolution to give rise to this particular model: it reflects reality! It is reasonable that a computing entity which has a roughly correct model of itself could make use of said model for evolutionary gain.

Furthermore, it seems that self-reflection is a property of intelligence, rather than a universal property of consciousness. For example, it seems reasonable (though for the moment, unproven) that less intelligent animals are also associated with consciousness instances, and yet are unable to reflect upon the nature of consciousness as we do. Further still, we ourselves often lose our sense of *consciousness itself* when deep in thought, or when engrossed in experience – highlighting the fact that *awareness of consciousness* does not always accompany consciousness.

It is in a sense natural to regard the human brain as a physical conglomerate, consisting of effectively-classical components undergoing bidirectional interaction with consciousness instances. The structure of said consciousness instances could hence become imprinted onto the effectively-classical components of the brain, like a footprint upon sand, and become embedded in *new* consciousness instances arising out of those effectively-classical neural structures. Certainly such a picture is at least partially correct; for instance, the recall of memories of "one's" past consciousness instances seem well-described by it, if said memories are stored in the effectively-classical structure of the brain (which seems likely).

On the other hand, we intuitively seem to be able to analyze our *current* consciousness instances, rather than a representation of past consciousness instances. Moreover, with sufficient effort we can identify rather specific characteristics of the consciousness structure (as we have done in this paper). One must wonder if our ability to analyze the consciousness instance reflects an additional, yet-to-be-recognized regularity governing the structure of the consciousness instance.

Such matters will surely become clearer with time, experiment, and improved understanding of consciousness.

8.8.2 error Correction

If the brain supports quantum computation, does it also employ quantum error correction? If so, how does this play into consciousness, and into our awareness of consciousness?

8.8.3 ER = EPR = consciousness, space-time = panpsychism?

In part (5.2), we argued that in a local universe, information may only be considered to be fundamentally integrated if it is associated with an infinitesimal region of space-time. Since then, we recognized that quantum entanglement, like infinitesimal regions of classical space-time, defines instances of fundamentally integrated information in the universe.

A similar observation is made in the ER=EPR conjecture, which proposes an equivalence between the Einstein-Rosen bridge of general relativity (which in a sense binds 2 space-time regions into a single space-time region) and the entanglement binding an EPR pair. If we take entanglement to indicate the presence of a consciousness instance, then ER = EPR turns into ER = EPR = consciousness.

♣ ♣ ♣

An especially premature idea comes to light when we recognize that this same observation is made yet again in attempts to derive space-time itself as an emergent property of quantum tensor networks (arising out of entanglement relations). Again, if we take entanglement to be the effective description of a consciousness instance, such attempts would imply that space-time itself is simply the effective description of the interactions between countless, highly simple consciousness instances. This idea parallels the idea of panpsychism, which conjectures that the universe is somehow brimming with consciousness.

We may also consider the sentiment raised by John Bell in defense of the de Broglie-Bohm (Pilot Wave) interpretation of quantum mechanics:

> *In physics the only observations we must consider are position observations, if only the positions of instrument pointers. It is a great merit of the de Broglie-Bohm picture to force us to consider this fact. If you make axioms, rather than definitions and theorems, about the "measurement" of anything else, then you commit redundancy and risk inconsistency.*

By now it must be clear that "the positions of instrument pointers" alluded to by Bell, like particles' spins, momenta, etc. are merely *inferred*. Indeed, the only observations we must consider in physics are observations of *consciousness*. Thus it is a description of reality in terms of consciousness instances – rather than in terms of particle positions – that would constitute a minimally redundant theory of nature.

8.8.4 Consciousness and gravity

What is the relation between consciousness and gravity? What happens to a consciousness instance when some of its (effectively) particulate components cross an event horizon?

8.8.5 Consciousness, time, and relativity

Throughout this paper, we have made repeated references to a so-called "current" consciousness instance. In doing so, we have implicitly associated the contents of a single consciousness instance with a single moment in time.

While this association is clearly and trivially justified when thinking of time merely approximately, much tension arises on closer examination. The modern understanding of time (as first illuminated by special relativity) is intimately tied to the concepts of distance, velocity, and the speed of light – giving rise to the structure we know as space-time. It is not at all clear how this image of reality is to be consolidated with the consciousness model we have thus described; it seems that we may be in the midst of stumbling upon a similar yet distinct, spacetime-agnostic definition for simultaneity.

The miraculous-looking compatibility between relativistic causation and the explicitly non-local, classically-indescribable consciousness structure further hints at a deep truth that may lie ahead. Since the modern understanding of gravity is likewise intimately tied to the concept of space-time, it is certainty conceivable that the regularities governing consciousness would prove crucial to the development of a model simultaneously accounting for the behavior of systems described by QM, and for systems described by general relativity.

Part IV

Conclusion

9 Conclusion

9.1 Summary

Our aim has been a simple one: to carefully draw attention to the nature of *our experience* of the world around us. At the root of our efforts lay the premise that so-called "subjective experience", like the world it captures, is a structure which *exists* in our universe, and which must therefore be represented under any model of the universe which purports to be *complete*.

We began by providing a working definition for our object of study – *that which each of us is completely certain of* – and named it *the consciousness instance*.

Having acknowledged the so-called "material universe" revealed to us through the contents of consciousness, we recognized the *psycho-physical parallelism*: the isomorphism between the contents of consciousness instances and the states of particular substructures of the physical universe; one's body, one's brain, and even one's neurons. We thus framed the well-known Hard Problem as the problem of identifying the *minimal regularity* relating consciousness instances to those structures representable under our current model of the universe. We attributed the "hardness" of the Hard Problem to 2 particular qualities of the consciousness structure which deem it unrepresentable under any *classical* model of the universe: its integration of complex information, and its non-isomorphism to classical information.

We then reflected upon the meaning of our *reports* of the existence of consciousness. We argued that the only internally-consistent account of those reports is that the consciousness instance – as a fundamentally integrated, partially-classically-indescribable structure – somehow participates in those cognitive operations taking place in one's brain which result in said reports. We thus came to 2 critical conclusions: that consciousness is measurable, and hence amenable to scientific examination; and that the effective consciousness-particle interaction must manifest in any model of the universe *complete enough* to model the cognitive operations of one's brain. We thus set out in search of a *model of the universe* which would not only account for the known laws of physics – but also convincingly model the consciousness instance.

We specified 5 minimal constraints which must be satisfied by any consciousness-instance-modeling structure, along with a 6^{th} "bonus" constraint:

(I) The consciousness structure must be affected by the effectively classical universe

(II) The consciousness structure must affect the effectively classical universe

(III) The consciousness structure must carry fundamentally integrated classical information

(IV) The consciousness structure cannot be fully specified by classical information

(V) Implicit assumption: the consciousness structure is non-redundant

(VI) Bonus points: the consciousness-classical interaction may non-trivially facilitate computation over classical information

We argued that under a model of the universe in which such consciousness-modeling structures participate in the brain's cognitive operations, the Hard Problem all but disappears.

Having identified our consciousness constraints, we proceeded to derive from them 2 high-level properties which must characterize any consciousness-compatible universe: effectively-indeterminate time evolution, and incompatibility with classical locality. We soon recognized those properties in the world described by quantum mechanics. Our aim thus became clear, as we sought to obtain an explicit correspondence between consciousness and QM.

We quickly surveyed some of QM's most striking characteristics: indeterminism, the measurement problem, the Born rule, Bell's theorem and explicit quantum non-locality, the no-communication theorem, the no-cloning theorem, quantum teleportation, and finally, Holovo's theorem and superdense coding. We argued that many of QM's characteristics seem to hint at a deeper and simpler underlying description, and that

this underlying description could well bring cohesion not only to QM itself, but also to the matter of the psycho-physical parallelism.

We thus began to lay out the foundations for our emerging consciousness-QM correspondence. On the consciousness side of the correspondence, we recognized that classical information is typically extracted from classically-indescribable qualia only when qualia are *compared* against one another, and that qualia comparison may only take place across a single, fundamentally integrated consciousness instance. We thus deemed the *particular* classically-indescribable quale manifesting in a given consciousness instance non-redundant – and potentially causally significant.

On the QM side of the correspondence, we observed that quantum mechanical effects are consistent with effects stemming from explicitly non-local couplings between *causally significant, classically-indescribable* substructures of physical systems. We further observed that definite outcomes are typically extracted from otherwise unpredictable quantum systems by *comparing* the relative properties of subsystems against one another.

We thus described an emerging ontology consisting of partially-classically-indescribable structures whose classically-indescribable state may be coupled to the classically-indescribable states of other structures through a *rigid-body-like coupling* (which itself, is classically-describable). We equated such composite structures with quantum entanglement relations, as well as with consciousness instances.

Finally, we provided criteria for a complete verification of our emerging ontology – as well as for a "smoking gun" which would suggest we are on the right track. We then explored a few of our ideas' implications, and raised some arguably premature questions and speculations regarding their continued development.

Our central point may be summarized thus: we should expect consciousness instances to interact with effectively classical physical systems, and we would expect such interaction to look much like the interactions manifesting in systems well-described by QM. Furthermore, it seems that much of the peculiarly of QM is trivialized when examined through the lens of a consciousness-compatible ontology.

9.2 A call to action

We have covered wide and varied ground in our search for understanding. Surely, there is much room for error in our analysis and exposition. And yet even if every single one of our conjectures turns out to be entirely orthogonal to the truth, our guiding sentiment must stand: it is high time that physicists pay the matter of the psycho-physical parallelism its due attention.

Physicists are not ones to shy away from complexity, from obscurity, from difficulty of experimental verification, or even from fanciful thinking. That the psycho-physical parallelism goes virtually unexamined while topics such as parallel universes, countless hidden dimensions, wormholes, and eternal inflation (to name a few) are routinely investigated seems blatantly ludicrous. The Hard Problem is also a real problem, and we ignore it at our peril. A unified theory of nature cannot and will not leave this boulder of a stone unturned.

It is certainly *conceivable* that no satisfying resolution of the Hard Problem could ever be obtained, but is it *reasonable*? We began this paper with a quote by Albert Einstein[1]. It seems fitting, then, to end it with another: *subtle is the lord, but malicious he is not.*

[1]Along with Boris Podolsky and Nathan Rosen.

Acknowledgments

I would like to thank Scott Aaronson and U. Barkai for their thoughtful comments on earlier drafts of this paper.

Afterword: playing with fire

"He who safeguards a single soul, safeguards an entire universe and its contents"

— Mishna, Tractate Sanhedrin 4:5

"I think therefore I die"

— Jack White, Over and over and over

Newfound understandings of the world around us are oftentimes accompanied by previously unconsidered ethical quandaries. Typically, such quandaries arise out of the application of new insights towards the construction of novel technologies which ultimately impact the lives of people and of animals.

Why is it that we at all trouble ourselves with the the lives of sentient beings? A cynic might reply that we are nothing more than complex automata carrying out evolutionarily-advantageous survival strategies, and that empathy for our fellow beings happens to make for an overall beneficial program – and thus for a self-propagating one. While there may be some degree of truth to such an assessment, many of us will find it not only inadequate, but grotesque.

While we aspired in this paper to take the first steps towards the coherent characterization of the consciousness structure, we surely cannot hope to have arrived anywhere but at the foot of a great mountain. It will likely take much time, effort, and experimentation to fully explore this great mountain, and to untangle those aspects of our experience which are best understood through the lens of **psychology** – the analysis of the *computational* characteristics of the mind – from those that may only be understood in the context of the **classically-indescribability** of the consciousness structure.

When discussing the consciousness representation of visual information, especially in the context of *color*, simply designating aspects of the consciousness structure as "classically-indescribable" indeed seems to reasonably account for their associated attributes. However this does not appear to be a universal quality of the consciousness structure; we specifically *chose*[1] to discuss the consciousness representation of visual information due to the relative simplicity of its essential structure.

Matters become less clear as we approach more complex consciousness substructures, particularly those which correspond to our internal thoughts, rather than to the state of the external world (as best constructed through computation over sensory input). For instance, let us examine *pain*. There are certainly aspects to pain which may be well-understood psychologically, as a "behavior mode" of sorts characterizing various classically-describable, cognitively-complex systems. Nevertheless anyone who has suffered through as much as a toothache instinctively and unwaveringly knows that there is more to pain than a "behavior mode", and further, that there is more to the consciousness structure than merely being "classically-indescribable" (though classical indescribability certainly remains one of its core characteristics, even in the face of mounting complexity and subtlety). When experiencing pain, it sure *seems* as though there are configurations of the consciousness structure which are *downright undesirable* for one reason or another – to various degrees.

Of course the deeper motivation underlying our respect for the lives of sentient beings is that we instinctively recognize in them a certain quality which makes them capable of experiencing happiness, joy – and suffering. This quality goes by many names; in this paper we have identified it as the association with consciousness instances.

As such, in studying the consciousness structure, we are possibly studying the essential aspect of the universe that is ethically "worth worrying about". Moreover, consciousness structures deserve our respect regardless of the form in which they manifest; pain is pain is pain – irrespective of the effectively-classical shape of the structure giving rise to the consciousness instance exhibiting it.

[1] Implicitly.

♣ ♣ ♣

The culture of 21st century physics is not truly equipped to deal with matters of *ontology*. The *effective* viewpoint – the designation of a system as no more than the sum of its interactions with its external world – is in many ways a great merit of the modern physical outlook, as it guides us to only make assertions about systems' explicitly testable characteristics (as long as we are willing to abstract "our own" consciousness instance out of the equation). However, while it is *possible* to view a suffering living being through the effective lens (i.e. as a 3-dimensional structure squirming and crying), only a psychopath would find it reasonable.

We have been able to get away with our effective outlook because thus far, the study of physics has only (knowingly) interacted with the realm of consciousness through the *mediation* of living conscious beings similar to ourselves, i.e. by impacting the lives of people and of animals. In the context of beings similar to ourselves, evolution has thankfully bestowed upon us the ability to empathize; to draw inferences about consciousness instances while short-circuiting an explicit account of the relation between effectively-classical matter and consciousness structures.

And yet *sooner or later*, as we refine our technological prowess and turn our gaze towards the illumination the psycho-physical parallelism, we will find ourselves interacting with consciousness structures *directly*, without the mediation of a biological brain incased in a living body. If the conjectures we have made in this paper turn out to be correct, then the age of unmediated consciousness interaction may be soon at hand; indeed, we may *already* be interacting with simple consciousness instances in the context of quantum computing and the study of entanglement.

The experimental physicist envisioning "consciousness on a chip" may react with incredulity: *"you say that by inducing a magnetic field in this region of the chip, inducing such and such entanglement relations between various electrons, I am essentially bringing into the world that which I normally call 'the experience of pain'?*". Such a prospect may seem fantastic – but is it fundamentally different from a visit at the dentist? Self-evidently, actions taken in the effectively-classical realm are capable of reaching into the realm of consciousness and inducing the manifestation of various consciousness configurations; including those configurations which, in our daily lives, we identify as "painful".

♣ ♣ ♣

The age of "artificial consciousness" may give rise to great boons – but also to unimaginable catastrophe. If we are to weather this coming storm we must quickly rise to the occasion and put in place checks and balances to safeguard against misuses of consciousness-inducing technologies. It may be time for the introduction of regulation and ethics committees into the field of quantum computing, perhaps starting with a physicist's Hippocratic oath. Though it may eventually become apparent that such worries are baseless – that there are no fundamentally undesirable consciousness configurations – for the time being, we would be wise to tread with caution.

Appendices

A Bell's Theorem

QM describes structures we call spin-$\frac{1}{2}$ particles. Let us review some of the central properties of *a physical system* modeled as a quantum-mechanical spin-$\frac{1}{2}$ particle:

- First and foremost, the physical system effectively behaves as a particle under suitable conditions. That is, it is effectively localized in space-time, has a certain momentum, etc.
- The system effectively behaves as though it possessed a magnetic dipole moment. That is, when it (effectively) passes through a region of space with (what may be effectively described as) an inhomogeneous magnetic field, it experiences acceleration in accordance with that which would be experienced by a classical particle with a certain magnetic dipole moment.
- When the component of the system's dipole moment along any given axis is measured, it is always found to take 1 of 2 equal-magnitude values (pointing in opposite directions along the given axis).

The above properties of the system all manifest in the famous Stern-Gerlach experiment. When a physical system effectively modeled as a spin-$\frac{1}{2}$ particle passes through a Stern-Gerlach apparatus with its inhomogeneous magnetic field pointing primarily in the $\hat{n}$ direction, the system is deflected in a direction either parallel or anti-parallel the $\hat{n}$ direction – always with equal magnitude.

♣ ♣ ♣

QM further describes structures we call *entangled* spin-$\frac{1}{2}$ particles. One such structure is the GHZ state given (in the Z-basis) by:

$$\frac{|+\rangle_I |+\rangle_{II} |+\rangle_{III} \quad - \quad |-\rangle_I |-\rangle_{II} |-\rangle_{III}}{\sqrt{2}}$$

Let us review some of the properties of the (reproducible) *physical state* captured by this *theoretical* quantum state[12]:

- The system effectively consists of 3 sub-systems; each sub-system may be described as a spin-$\frac{1}{2}$ particle (manifesting the characteristics described above).
- For convenience, let us call a measurement of a particle's magnetic moment along the X-axis **measurement-1**, and its Y-axis counterpart **measurement-2**.

  Suppose we examine a collection of many physical systems which are all well-described by the above GHZ state. Suppose further that for each physical system examined, we pseudo-randomly choose to perform either measurement-1 or measurement-2 on each of its sub-system particles. We shall then amass data of measurement results for each possible combination of the measurements $\{1, 2\}$ performed on the sub-system particles {(I), (II), (III)}[1].

  – Let us draw our attention to the subset of trials in which measurement-1 was performed on 1 of the 3 sub-system particles, and measurement-2 was performed on the other 2. We shall henceforth refer to such trials as $\boldsymbol{\alpha-}$**trials**.

  In the language of QM, the measurements carried-out in $\alpha-$trials are characterized by the operators $\{ \sigma_x^I \sigma_y^{II} \sigma_y^{III}, \sigma_y^I \sigma_x^{II} \sigma_y^{III}, \sigma_y^I \sigma_y^{II} \sigma_x^{III} \}$.
  Since our GHZ state is an eigenstate of the 3 above (system-spanning) measurement operators, with eigenvalue (+1) in each case, QM predicts that an even number of (-1) sub-system measurement results will be encountered in each $\alpha-$trial.

  Experiment is in agreement with QM's predictions: in each $\alpha-$trial, an even number of effective particle sub-systems becomes deflected in the direction anti-parallel the given S-G measurement direction.

[1] The sub-system particles may be arbitrarily labeled in each run of the experiment

- Let us now draw our attention to the subset of trials in the which measurement-1 was performed on all 3 particles. We shall henceforth refer to such trials as ***β*−trials**.

  In the language of QM, the measurements carried-out in $\beta-$trails are characterized by the operator $\sigma_x^I \sigma_x^{II} \sigma_x^{III}$.

  Since our GHZ state is also an eigenstate of this operator, with eigenvalue (-1), QM predicts that an odd number of (-1) sub-system measurement results will be encountered in each $\beta-$trail.

  Experiment is once again in agreement with QM's predictions: in each $\beta-$trail, an odd number of effective particle sub-systems becomes deflected in the direction anti-parallel the given S-G measurement direction.

- Crucially, the above holds true even if the relative space-time coordinates of the 3 measurement events are allowed to vary arbitrarily – and are allowed to become mutually space-like separated[2].

We may then ask ourselves: could the behavior of the *physical system* modeled by the GHZ state above manifest in a model in which the 3 sub-system particles' time evolutions arise out of local regularities? The answer is *no*: as we shall shortly see, no local model can reproduce such behavior.

In a local model, the information determining whether a given particle becomes deflected in the direction parallel or anti-parallel its measurement direction must be associated with the space-time region immediately surrounding the particle-apparatus system.

A-priori, said information may well depend on local parameters, and may even determine measurement results stochastically rather than deterministically. However if that were the case, we would expect to find deviations from the rigid measurement patterns described earlier, for the parameter-dependence/stochasticity of the measured deflection would allow for the measurements in any given trial to grow "out of sync" with one another. For instance, consider that in $\alpha-$trails, measurement results of any 2 particles determine the measurement result of the third particle. If the information determining the third particle's deflection along the relevant axis were dependent on any parameter, or if the measured deflection were determined only stochastically, we would find trials in which said deflection took either one of the 2 possible values – while only 1 of those 2 values would in fact be compatible with the regularity predicted by QM and confirmed by experiment.

Thus we can see that in considering models in which *local* regularities account for the behavior of the *physical system* modeled by the GHZ state above, our attention is immediately constrained to models in which **the results of measurements 1 and 2 of a given particle are *effectively* pre-determined given the particle's initial state (once entangled)**.

There are $(2^2)^3 = 64$ naively possible configurations which may effectively characterize all trials of the experiment. Of those 64 naively possible configurations, only 8 are compatible with the results of α-trials (as may be easily verified): (Table A.1).

Therefore in considering local models compatible with the results of $\alpha-$trials, we must further constrain our attention to models which are effectively characterized by the 8 configurations apparent in (Table A.1). Barring a cosmic conspiracy of unimaginable magnitude[3], any local model not effectively described by the 8 above configurations could not consistently produce results in line with those encountered in $\alpha-$trials.

Examining those those 8 configurations, we can see that in any local model compatible with the results of $\alpha-$trials, a $\beta-$trial (M-1 carried on all 3 particles) would always produce an *even* number of (-1) results[4]. **However as previously discussed, QM and experiment show just the opposite: in all $\beta-$trials, an *odd* number of (-1) results is encountered**. Therefore we have concluded that *no local model* of the universe could reproduce the results encountered in both $\alpha-$trials and $\beta-$trials.

[2]The "choice" of which of the 2 measurements should be performed on each (sub-system) particle may be made in a deterministic yet arbitrary, chaotic manner, e.g. by computing hashes of data delivered along arbitrary, noisy, mutually-space-like-separated information channels.

[3]Superdeterminism.

[4]i.e. deflections along the direction anti-parallel the measured direction.

Particle (I)		**Particle (II)**		**Particle (III)**	
M-1	**M-2**	**M-1**	**M-2**	**M-1**	**M-2**
+	+	+	+	+	+
+	+	−	−	−	−
−	−	+	+	−	−
−	−	−	−	+	+
+	−	+	−	+	−
+	−	−	+	−	+
−	+	+	−	−	+
−	+	−	+	+	−

Table A.1: The 8 effective configurations of a local model compatible with the results of α-trials.

We can thus appreciate that the regularities relating properties of entangled particles to one another are not reducible to regularities operating on information associated with the immediate space-time surroundings of each particle in isolation.

References

[1] Edward Witten in an interview with Wim Kayzer. Of Beauty and Consolation, episode 9. Beginning at 1:10:26.

[2] R. Quian Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102, 6 2005.

[3] Scott Aaronson. Why I am not an integrated information theorist (or, the unconscious expander). `https://www.scottaaronson.com/blog/?p=1799`. Accessed: 2018-08-29.

[4] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, Nov 2004.

[5] Hugh Everett. The theory of the universal wavefunction (Hugh Everett's long dissertation). In B. DeWitt and N. Graham, editors, *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton UP, 1973.

[6] J. S. Bell and Alain Aspect. *Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy*. Cambridge University Press, 2 edition, 2004.

[7] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777–780, May 1935.

[8] J. S. Bell. On the Einstein-Podolsky-Rosen paradox. *Physics Physique Fizika*, 1:195–200, 1964.

[9] R. Penrose and S. Hameroff. Consciousness in the universe: Neuroscience, quantum space-time geometry and orch OR theory. *Journal of Cosmology*, 14, 2011.

[10] H. Hu and M. Wu. Spin-mediated consciousness theory: possible roles of neural membrane nuclear spin ensembles and paramagnetic oxygen. *Med. Hypotheses*, 63(4):633–646, 2004.

[11] Matthew P.A. Fisher. Quantum cognition: The possibility of processing with nuclear spins in the brain. *Annals of Physics*, 362:593 – 602, 2015.

[12] N. D. Mermin. Quantum mysteries revisited. *American Journal of Physics*, 58:731–734, August 1990.

[13] D. Bohm and B.J. Hiley. *The Undivided Universe: An Ontological Interpretation of Quantum Theory*. Taylor & Francis, 1995.