Could the physical world be emergent instead of fundamental, and why should we ask? (short version)

Markus P. Müller

Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna
Perimeter Institute for Theoretical Physics, Waterloo, ON N2L 2Y5, Canada

December 5, 2017

Abstract

In physics, there is the prevailing intuition that we are part of a unique external world, and that the goal of physics is to understand and describe this world. This assumption of the fundamentality of objective reality is often seen as a major prerequisite of any kind of scientific reasoning, delineating science from pseudoscience, and explaining why successful empirical science is possible in the first place. However, here I argue that we should consider relaxing this assumption in a specific way in some contexts. Namely, there is a collection of open questions in and around physics that can arguably be addressed in a substantially more consistent and rigorous way if we consider the possibility that the first-person perspective is ultimately more fundamental than our usual notion of external world. These are questions like *which probabilities should an observer assign to future experiences if she is told that she will be simulated on a computer? How should we think of cosmology's Boltzmann brain problem or assign probabilities to properties of 'possible worlds'? What can we learn from the fact that measurements in quantum theory seem to do more than just reveal preexisting properties? Why are there simple computable laws of physics in the first place?* This note summarizes a longer companion paper which constructs a mathematically rigorous theory along those lines, suggesting a simple and unified framework (rooted in algorithmic information theory) to address questions like those above. It is *not* meant as a 'theory of everything' (in fact, it predicts its own limitations), but it shows how a notion of objective external world, looking very much like our own, can provably emerge from a starting point in which the first-person perspective is primary, without apriori assumptions on the existence of 'laws' or a 'physical world'. While the ideas here are perfectly compatible with physics as we know it, they imply some quite surprising predictions and suggest that we may want to substantially revise the way we think about some foundational questions.

Contents

1 Introduction

> *"The hypothesis that there is an external world,*
> *not dependent on human minds, made of* something,
> *is so obviously useful and so strongly confirmed*
> *by experience down through the ages that we can say*
> *without exaggeration that it is better confirmed*
> *than any other empirical hypothesis. So useful is the posit*
> *that it is almost impossible for anyone except a*
> *madman or professional metaphysician to comprehend*
> *a reason for doubting it."* (M. Gardner [2])

This paper takes the perspective of the madman: only if we doubt the obvious will we be able to approach some longstanding open problems in physics and beyond, including the question *why* we see something like an external world at all.

This is a summary of a longer companion paper [1] which contains all the mathematical proofs referred to below and a much more detailed argumentation.

2 An unfamiliar but simple theory

2.1 The postulates: observers, probability, and Solomonoff induction

The starting point of this proposal is the idea that the problems mentioned in the abstract motivate a departure from some aspects of the traditional way that we tend to think about the world. *Traditionally*, a physical theory presupposes the existence of an objective material external world that evolves according to certain physical laws. Our theories about these laws are tested by calculating their predictions and by comparing them with the observations that we actually make. Since the discovery of quantum mechanics, we think that these predictions are probabilistic at best, and in principle all of the form

$$\mathbf{P}(\text{next observations}\,|\,\text{previous observations}). \tag{1}$$

For example, in a laboratory experiment, "previous observations" includes all our knowledge about the experimental setup and data we have acquired earlier; the "next observations" correspond to possible outcomes of the experiment. Crucially, we traditionally view the probabilities in (1) as being *derived* from (or secondary to) an objective external world[1]; either they arise because we are agents inside that objective universe who have only limited knowledge, or because the postulates of quantum theory claim directly that we should assign these probabilities as a consequence of the world's quantum state[2].

It turns out that several important conceptual problems in physics and beyond can be phrased in terms of the probabilities (1), and challenge the traditional view described above:

- *Quantum mechanics.* According to Bell's Theorem, naive versions of realism (roughly, the idea that measurement outcomes are always predetermined before the measurement is performed) are inconsistent with other important principles of physics (like locality). This has led to the slogan that "unperformed experiments have no results" [3], and to decades of discussions about how to interpret the counterintuitive formalism of quantum mechanics [4]. Substantial effort has been invested in tackling the question *"where the probabilities in (1) come from"*, without final consensus.

[1] Unless the probabilities are zero or one, in which case we often think of the corresponding propositions as reflecting actual properties of the world.

[2] Some readers will have strong views on the interpretation of probability (and/or quantum theory), but I suggest remaining agnostic on this question for the time being. How "probability" is supposed to be understood in this paper will become clear from the context and/or will have different possible interpretations. More details on this are given in [1].

- *Cosmology.* If we are observers in a really "big" universe (for example, a world undergoing eternal inflation), then the question arises regarding which probabilities of the form (1) we should actually assign to our own future observations. There are deep and surprising problems that arise in this context, for example the infamous *Boltzmann brain problem* [5, 6] (claiming that we should assign high probability to being only a short-lived thermal fluctuation in some cosmological models [7]), or, more broadly speaking, the *measure problem* of cosmology [8].
- *Artificial Intelligence / Philosophy of Mind* [9]. Even though it sounds like science fiction at the time of writing, current scientific progress suggests that we will soon live in a world where novel technologies present us with severe philosophical dilemmas. As one extreme and illustrative example, think of simulating the brain of a terminally ill person (after her death) on a computer [10]. Would this be a valuable endeavor? Would the person "feel like being" in the computer simulation, or would it have no effect on her first-person perspective whatsoever? Questions of this form can (at least in principle) be recast in terms of the conditional probabilities (1): what is the probability that the person is going to observe the simulated state of mind, given what she has observed in the past?

In this paper and in [1], I suggest to address aspects of all these questions in a unified way, by taking a radically unconventional perspective and by asking: *what if the probabilities in (1) are actually fundamental, and physics as we know it is an emergent phenomenon?*

As implausible as this may at first sound, there is a clear-cut technical starting point, namely algorithmic information theory [11, 12] and "Solomonoff induction" (SI). In a nutshell, SI suggests that any observer who has made previous observations x should assign conditional algorithmic probability $\mathbf{P}(y|x)$ to future observations y. The quantity that we call $\mathbf{P}$ here is defined in algorithmic information theory as a "universal apriori probability". It is a normalized version of a "universal enumerable semimeasure", corresponding to the probability that a randomly chosen (say, by a fair coin toss) computer program for a universal monotone Turing machine will output y after it has output x. Algorithmic probability is a mathematically natural quantity that can be defined in many equivalent ways and that has a multitude of applications in computer science; see [11] for the detailed definition and [1, Section 5] for an accessible overview.

This prescription yields a method of inference that is guaranteed to yield correct answers in the following sense: if the observations arise from a stochastic process which is computable (i.e. that it has a finite effective description for a probabilistic Turing machine), then the probabilities assigned by SI are guaranteed to converge to the distribution that is given by the process. For example, suppose that an observer sees one bit (zero or one) after the other, and she does not know that these bits are generated by a deterministic (stochastic) process that always outputs "1". After having seen n ones, SI suggests to assign the probability $\mathbf{P}(0|\underbrace{11\ldots1}_{n})$ to seeing a zero next, and it turns out that this goes to zero (roughly like $1/n$) as it should [12]. But then, it follows in particular that SI can be used for successful prediction of observations *in our physical world*, since the laws of physics as we know them are computable (as stated by a physical version of the Church-Turing thesis[3]). This is the reason why SI (or rather practically implementable versions of it) is considered applicable in artificial intelligence. In a nutshell, we have the following

Observation: Whenever our physical theories give us a concrete value of $\mathbf{P}_{\rm phys}(y|x)$, this prediction will agree with Solomonoff induction's $\mathbf{P}(y|x)$ — at least asymptotically, after many observations.

[3] See [1, Section 2] for more details and references, and [13] for a definition of the physical Church-Turing thesis. In a nutshell, the version that I am using here claims that there is an algorithm that yields a description of the *probabilities of outcomes*, given the description of any experiment. It does *not* mean that the actual outcomes can be predicted or computed. For example, an infinitely repeated fair coin toss is computable according to this definition (since it is a stochastic process with a finite description), but the actual outcome sequence will almost surely not be computable. The laws of physics (at least in the approximate form that we know them) seem to be computable in this sense.

In other words, we can in principle make correct probabilistic predictions *by applying SI alone*, without any direct reference to physical theories. This suggests two possible routes of exploration.

First, it suggests that we can *use SI as a pragmatic "rule of thumb" whenever physics itself does not give us any obvious probabilistic predictions.* For example, in the cases of cosmology or brain emulation sketched above, it is not clear how our physical theories would allow us to assign conditional probabilities of the form (1); in fact, some philosophers would argue that physics is in principle unable to do so in some cases. In these cases, we can instead use $\mathbf{P}$ directly for prediction. Ignoring the practical difficulties of doing so, this yields a pragmatic method of inference, motivated by the observation above and also by further considerations like Ockham's razor (since $\mathbf{P}(x)$ is larger for simpler x, i.e. for those that have a shorter description). This will give us predictions in realms where physics in itself does not, and is in principle guaranteed to be compatible with physics in regimes where SI and physics are both applicable.

Second, it suggests a much more ambitious idea: *what if there is only a single fundamental law, namely that algorithmic probability determines future observations?* Could it be that the physical laws and regularities that we observe (including the appearance of an objective external world) are simply *consequences* of (Solomonoff) induction? If so, this would imply a worldview that is quite different from the standard one, more similar to Wheeler's idea of "Law Without Law" [14].

Before examining the conceptual basis of this in more detail, let us turn to the technical question of how one can obtain a concrete theory from this idea. In this summary paper, I will drop most mathematical details for reasons of brevity; all these details and much more thorough discussions can be found in [1].

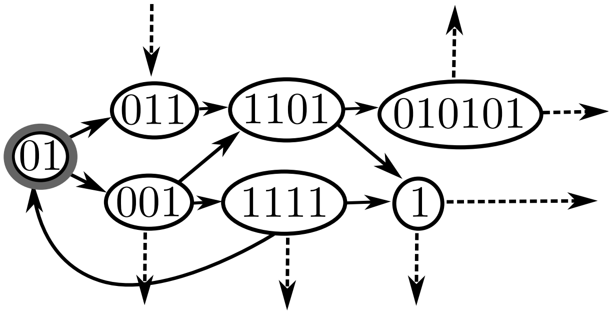

Figure 1: Schematic illustration of an observer graph A, which is a computable directed rooted graph on the finite binary strings (here the root is $\Lambda = 01$). Any path through the graph (starting at its root) will be called an "A-history". For any vertex x, we denote by $A(x)$ the set of all vertices y such that there is an arrow from x to y. For example, in this figure, we have $A(01) = \{011, 001\}$. An *A-measure* $\mu(\bullet; A)$ is a probability measure on the A-histories, i.e. $\mu(\Lambda; A) = 1$ for the root Λ, and $\sum_{y \in A(x_n)} \mu(x_1, \ldots, x_n, y; A) = \mu(x_1, \ldots, x_n; A)$. In particular, $\mathbf{P}(\bullet; A)$ as defined below is an A-measure. Standard algorithmic information theory [11] is mostly restricted to the special case of the tree graph where Λ is the empty string ε, and $A(x) = \{x0, x1\}$.

We start by defining the notion of an observer. Note that the purpose of this definition is neither to capture what we colloquially mean by an observer, nor to decide once and for all how we should think of an observer, but rather to abstract the most important features of it to allow for a mathematically sound theory. We will do this by introducing the notion of an "observer graph", which is a (computable, rooted) directed graph over the finite binary strings, $\{0,1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, \ldots\}$. This captures the following idea. Any observer (for concreteness, think of a human being or animal for now) will, at some moment, contain information that encodes everything that she sees, knows and remembers at that moment, described by some (usually very long) binary string $x \in \{0,1\}^*$. Naively, think of encoding the full content of the brain into a long string of zeroes and ones. In the following, I will interchangeably use the words "experience", "observation" and "state of the observer" for such a string, emphasizing different aspects of the interpretation, but without commitment to any details of the interpretation (in particular, no direct relation to "consciousness" etc. is claimed in using these words). For example, this string could describe the experience of a bat, flying inside a cave towards a turning point where it cannot see what is coming next. Then, one moment later, there will be another string $y \in \{0,1\}^*$ that describes the observer's next experience. In general, there are many *possible* next strings y; for example, the bat might see that the cave just goes on, much further, in the same way as before; or she may be very surprised to find the cave's end, since part of it has collapsed since she had been there the last time.

We will formalize this by having the string x as a vertex of a graph, and (at least) two arrows pointing away from x, to the two possible next strings y. As a result, we get a graph as in Figure 1 (just think of the strings as being typically much longer). Every vertex (binary string) describes a conceivable momentary experience of the observer, while the outgoing arrows point to the next possible experiences. Note that "possibility" is here not defined with reference to any laws of physics (there are none at this point), but should rather be understood as "subjectively legitimate successor experience". For example, in the case of the bat, another possible next experience y (following x as described above) would be to see, suddenly and surprisingly, a huge massive rock made of gold materialize in the cave where she is flying. This would correspond to some arrow in the observer graph, even if our idea of the bat as embedded in a physical world would make us expect that this experience will be physically disallowed. On the other hand, there would *not* be an arrow to a string that describes the experience of, say, Donald Trump on a state visit in Austria, looking up to the mountains (and having all his usual memories). This would not correspond to a subjectively legitimate successor experience of the poor bat.[4]

Of course, the handwaving argumentation above raises all kinds of questions, like: *how* should we encode some observer's state into a binary string? Or, *which* transitions (arrows) are concretely allowed, and which are not? The point is, however, that we do not need to answer these questions in order to write down the postulates of our theory. And then, once the theory is in place, we will actually see how to answer those questions. For example, regarding the former question, it will turn out that the choice of encoding is irrelevant, since predictions of the theory will be "covariant" with respect to the choice of encoding, in a somewhat similar way as equations of General Relativity are with respect to a choice of coordinate system. Moreover, we never really have to perform or compute an actual encoding, since most predictions will follow from the existence of such an encoding and not from its exact form. Regarding the latter question, it will turn out that we can (and should) actually allow *all* transitions, and have the complete graph as our observer graph. This is because algorithmic probability by itself will make sure that "subjectively non-legitimate successor experiences" (like the bat becoming Trump) are extremely unlikely.

Postulates of the theory (see also Figure 1):

1. Every observer is described by an observer graph A as defined above; the sequence of observations that the observer successively experiences corresponds to the list of binary strings in an A-history.

   Remark. At the end of [1], we amend this postulate by demanding that A has to be the *complete* graph on all strings[5]; we keep the more general formulation initially for convenience.

2. After having experienced a finite A-history $\mathbf{x} = (x_1, \ldots, x_n)$, the observer will subsequently experience one of the strings $y \in A(x_n)$ at random. The probability of every $y \in A(x_n)$ is given by the conditional algorithmic[6] probability $\mathbf{P}(y|\mathbf{x}; A)$.

[4] This concept is similar to Parfit's "Relation R" [15]. Nevertheless, the exact definition of this relation turns out to be irrelevant for the theory presented here; in fact, we will soon see that we can drop it altogether.

[5] This means that there are no apriori restrictions on the possible transitions whatsoever, which simplifies the theory. However, for measure-theoretic reasons, the graph will still have to be *rooted*, i.e. we have to specify the initial state in which the observer starts. See also the discussion in Subsection 2.4.

[6] Conditional algorithmic probability $\mathbf{P}(y|\mathbf{x})$ is usually only defined if y is a bit (zero or one), and $\mathbf{x}$ is a single string of bits. Thus, we have to extend the definition to the case that y is a bit string and $\mathbf{x}$ is a sequence of strings as above (and this has been done in [1]). Technically, it is a generalization of the universal enumerable semimeasures on continuous sample spaces of [11], relying on a straightforward generalization of the monotone Turing machine that I call a "graph machine".

This is it — there are no further postulates or assumptions. All other aspects of physics that we, as human observers, experience and cherish (including the appearance of an external world, and intersubjective agreement between different observers) are not postulated, but rather are expected to emerge as provable consequences. In this sense, this is a theory which is arguably as simple as possible: it only postulates that we make a sequence of observations (which is the only thing we definitely know), and it makes a statement about the propensity of the possible observations. These are the minimal ingredients of any physical theory.

Before we discuss how well-known aspects of physics follow from these postulates, we have to briefly discuss one problem that readers with background in theoretical computer science will immediately acknowledge. By definition, $\mathbf{P}(y|\mathbf{x};A)=\mathbf{P}(\mathbf{x},y;A)/\mathbf{P}(\mathbf{x};A)$, and $\mathbf{P}(\mathbf{x};A)\equiv\mathbf{P}_U(\mathbf{x};A)$ is defined as the probability that a universal computer U outputs A-history $\mathbf{x}$ if it is given a random program as input, together with a description of A. *But there are infinitely many different universal computers U,* and so which computer U should we choose as our reference to define the probabilities? This question has been recognized as an important problem in artificial intelligence [16] and does not have an easy solution [17]. However, it can be shown with some effort [1, Sec. 6.1] that the theory that follows from the postulates above is invariant with respect to the choice of U (since the resulting transition probabilities are "covariant" [1, Lemma 5.15] with respect to this choice), as long as U is chosen from a specific infinite subset of universal computers.

2.2 Three roads to the same theory

So far, we have motivated our two postulates by the observation (framed on page 2) that Solomonoff induction's $\mathbf{P}(y|\mathbf{x})$ will give predictions which are asymptotically identical to those of our physical theories, suggesting that we can in principle (but not so easily in practice, of course) replace our laws of physics by instead postulating the correctness of SI. But we can also use this observation to motivate the postulates by a kind of **structural argument** [18]: if there is *one* canonical mathematical structure that gives us a notion of probability, or propensity, then we do not need to introduce an additional structure (another theory of physics) to explain these probabilities, since this would amount to logical overdetermination in some sense. This kind of thinking is ubiquitous in theoretical physics, as explained in [1, Section 7]. Here it suggests that we interpret the "physical world" as a "propensity structure" (i.e. something that determines which observations are more likely than others), and identifies algorithmic probability as a canonically distinguished choice of such a structure.

It is encouraging to see that there are several other routes of argumentation that lead to essentially the same postulates. One of them is the idea to explore what would happen in a **world "without any laws of nature"**[7]. As I argue in [1, Section 7], one approach to formalize this intuition is to demand that "there should be no preference of any given choice of laws of nature over any other possible choice", and chances of what happens (or rather what we observe) should be determined simply by the structure of mathematics itself (since it represents the collection of all logically consistent possibilities). The attempt to formalize the notion of a "completely random choice of mathematical structure" along these lines will arguably again lead to (some version of) algorithmic probability.

[7] Note that the idea of "absence of physical laws" has been discussed in several contexts before, for example in Wheeler's work and in QBism [19, 20, 4] (formerly known as Quantum Bayesianism). While the ideas here have a lot of overlap with the general worldview of QBism, they differ in particular in one point of view: namely, they emphasize that it is *so damn difficult* to explicate in detail what it means to claim that the world is "lawless", or that "the universe as a whole is still under construction" [21]. Historical "paradoxes" of probability like e.g. the Bertrand paradox should be a warning for us and an incentive to be precise when we make these statements. QBists can get around the general demand for precision in this context by simply saying that the empirical representatives of our theories (quantum states) have no "ontic hold" on the world [22], thus avoiding the task to *formalize* what they mean by saying that the world "kicks back" in an unpredictable way. My claim is that we currently know of only one way to formalize this idea, namely in terms of probability theory which is not interpreted in a subjective Bayesian way: by saying that some things *are in fact* more likely than others. Within this framework, we can proceed in a mathematical/structural way by interpreting "lawlessness" as "absence of a distinguished structure", which leads to large parts of the ideas presented here. But I invite everyone to come up with a better idea to formalize this intuition.

A third route is comparable with some ideas that cosmologists have cooked up, see e.g. Aguirre and Tegmark [23]. Let us for a moment imagine the multitude of all conceivable, logically consistent worlds or "universes" (note that the theory as constructed here does not claim that we have to think about the world like this). According to the usual picture of observers (supplemented for the moment by this multiverse picture), we would intuitively say that every observer is part of some universe. However, to every observer, there are infinitely many copies, subjectively indistinguishable, embedded in infinitely many (sometimes only slightly) different universes (or sometimes even several copies in one universe as in [23]). Suppose that there is *no* apriori notion of one universe being "real" and all others being "not real". Then, if an observer makes new observations that give her more indexical information, i.e. tell her that she cannot be in universe A, but can be in universe B (while both were consistent with her observations the moment before), then this should manifest itself for her as the outcome of a random experiment. In other words, she should see a **statistical mixture of all "universes" consistent with her previous observations**. It turns out [1, Section 7] that algorithmic probability can indeed be interpreted as a statistical mixture of all computable deterministic worlds.[8] This gives fans of a multiverse-like worldview an independent reason to adopt the two postulates above. However, these postulates do not claim the existence of a multitude of worlds, and the multiverse picture is just one possible interpretation among many.

2.3 So how do we get physics from this?

What would observers see if the two postulates were true, and no other assumptions (like the existence of an external world etc.) were made?

According to Postulate 2, an observer's experiences are determined by algorithmic probability $\mathbf{P}$. Even though $\mathbf{P}$ is a mathematically "natural" quantity that appears in many different contexts and can be defined in many different ways, it is in some other sense quite "complex" and irregular; for example, it is noncomputable. At first sight, this should imply that observers make very irregular and unpredictable observations. A second thought, however, paints a quite different picture: after all, algorithmic probability $\mathbf{P}$ favors compressibility by giving higher weight to histories $\mathbf{x} = (x_1, \ldots, x_n)$ that have a shorter description. A first result [1, Theorem 8.6] shows that this has an interesting consequence described in the following theorem. It uses the notion of a "computable test", which is a computable function f that maps A-histories $\mathbf{x} = (x_1, \ldots, x_n)$ to single bits (0 or 1), interpreted as the result of a "yes-no-question". For every such test, we demand in addition that it is "open", i.e. that it can yield both"yes" and "no" for future observations — formally, for every A-history $\mathbf{x} = (x_1, \ldots, x_n)$, we demand that there are $x_{n+1}, x'_{n+1} \in A(x_n)$ such that $f(x_1, \ldots, x_n, x_{n+1}) = 0$ and $f(x_1, \ldots, x_n, x'_{n+1}) = 1$. Then we get the following.

Theorem 2.1 (Principle of persistent regularities)**.** *Let A be a dead-end free observer graph, and f an open computable test. Suppose that f has given the answer "yes" to all observations in the past; then it will give the answer "yes" with high probability in the future. In more detail, for $b \in \{0,1\}$ define*

$$p(b|1^n) := \mathbf{P}\left(f(\mathbf{x}_1^{n+1}) = b \mid f(\mathbf{x}_1^1) = 1, f(\mathbf{x}_1^2) = 1, \ldots, f(\mathbf{x}_1^n) = 1\right),$$

where $\mathbf{x}_1^m := (x_1, x_2, \ldots, x_m)$. Then $p(1|1^n) > 1 - \frac{1}{n}$ for all but finitely many n. Moreover, the probability that $f(\mathbf{x}_1^n) = 1$ for all n is non-zero.

What this says is that regularities tend to stabilize themselves: if a computable regularity has been present (maybe by pure chance) for long enough, then it will tend to persist in the future. In some sense, observers will "catch lawlike regularities" like they would catch a cold — not as a consequence of some external "laws of nature", but simply due to the properties of algorithmic probability.

With this principle in place, let us discuss how the Boltzmann brain problem is automatically resolved. Suppose our observer (let's call her Bambi) is currently in a state in which she remembers having lived a

[8] This is not inconsistent with quantum theory as we will discuss in Subsection 2.5.

rich life full of experiences in a standard, low-entropic planet-like environment (and she has been like this in the past). Within the standard cosmological picture of our world, there is a possibility that Bambi, or rather her brain with all her memories, has just now appeared as a highly improbable thermal fluctuation, surrounded by a soup of thermal gas. In the next moment, this could mean that she makes a very strange and unexpected experience (say, heavy pain due to gas hitting her synapses). Let us call this a "BB-experience".

How probable is a BB-experience? If our universe is very large (say, due to eternal inflation), then naive counting may in some cosmological models suggest that a BB-experience is far more likely than our actual standard experience [5, 6]. However, according to our theory, Bambi's subjective experience is determined by algorithmic probability **P** as in Postulate 2 (and *not* by counting frequencies of events in some world), leading to the principle of persistent regularities above. But this principle says that *if the description of having evolved in a standard way on a planet worked very well in the past, it will probably persist in the future.* This is because we can always cook up an open computable test that asks whether Bambi's experience is typical for a planet-like environment or not[9]. Thus, a BB-experience is unlikely. In a nutshell, the reason is that a BB-experience has much higher conditional Kolmogorov complexity (given past experiences) than a standard experience.

The principle above is only referring to single computable tests. But what if we have several different computable tests, do the answers all fit together into some coherent overall lawlike behavior? The next theorem shows that there is indeed a tendency for this to happen:

Theorem 2.2. *Let A be a dead-end free observer graph, and μ a computable A-measure. Then*

$$\mathbf{P}\left\{\mathbf{P}(y|x_1,\ldots,x_n;A) \overset{n\to\infty}{\longrightarrow} \mu(y|x_1,\ldots,x_n;A)\right\} \geq 2^{-K(\mu;A)};$$

that is, with probability at least $2^{-K(\mu;A)}$ (which is large if and only if μ is simple), the actual transition probability $\mathbf{P}$ will in the long run converge[10] *to the computable measure μ.*

That is, asymptotically (i.e. after n observations, where n is large), observers will see that their probabilities of future observations are very well described by another probability measure μ. While **P** itself is noncomputable, μ instead *will* be computable and thus be much better suited for actual prediction. Moreover, with high probability, the Kolmogorov complexity $K(\mu;A)$ of the measure μ will be small. This means that μ will probably be simple in a very specific sense:

Simplicity of μ: *there exists a short computer program (of length $K(\mu;A)$) which makes the universal reference machine U do the following. If the input bits are chosen uniformly at random (as if by a fair coin toss), the machine produces a random output history distributed according to μ. This means that μ, as a stochastic process, has a short description; but it does* not *mean that the actual sequence of outputs is simple in the sense of having a short description.*

For example, let A be the observer graph that has just two vertices, 0 (which is also the root) and 1, and transitions between both are possible. Then one possible program for the universal machine would be to first output zero, and then to sequentially move the input bits to the output tape. This would be a very short program, and the measure μ that it generates is the random coin toss. Hence $K(\mu;A)$ would be very small, but the Kolmogorov complexity of the first m outputs would typically be about m, i.e. maximal.

So suppose that the event above indeed happens[11], and **P** gets close to μ for large n. Then observers may say the following: *"It seems that what happens is not completely deterministic, and sometimes pretty*

[9] While there are many conceivable ways to do this, we should construct a test that is as simple as possible (in the sense of description length/Kolmogorov complexity), since the convergence $p(1|1^n) \to 1$ will happen faster for simpler tests f.

[10] In more formal detail, the difference between $\mathbf{P}(\bullet|\mathbf{x_1^n};A)$ and $\mu(\bullet|\mathbf{x}_1^n;A)$ converges to zero in Hellinger distance.

[11] I do not currently know whether convergence to some computable measure μ must happen with probability one, i.e. whether $\mathbf{P}\{\exists\mu \text{ computable} : \mathbf{P}(y|x_1,\ldots,x_n;A) \overset{n\to\infty}{\longrightarrow} \mu(y|x_1,\ldots,x_n;A)\} = 1$. But even if this probability is strictly less than one, we can still hope that *some parts of the future observations y* will asymptotically be governed by computable laws, or that we have a weaker form of convergence in the sense that **P** converges to μ on all computable statistical tests, or something like this. The principle of persistent regularities motivates to consider the asymptotic emergence of computable regularities as a generic

complex... there seems to be intrinsic randomness in the world. However, this randomness seems to be governed by simple probabilistic "laws of nature" that have a very short description. Wow, that's interesting!"

However, these "laws of nature" μ are of a quite unusual form in this formulation: they say how likely *certain observations* are, *not* how likely certain events occur in the external world — there is not yet any notion of "external world". Yet, we *do* get an **emergent notion of external world**: as explained above, $K(\mu; A)$ being small means that there is a simple computational process that generates outputs distributed according to μ. This computational process has all the characteristics that we normally attribute to an "external world": it contains the observer (namely, the observations correspond to the outputs of the process), it evolves in time according to simple computational laws (since it is an algorithm), and most of it is not directly accessible to the observer. However, those parts of the computational process in addition to the outputs (that is, those parts which are not directly observed — say, the working memory) are correlated to future observations. Therefore, it makes sense for the observer to model and guess the state of these additional elements. This is analogous to a human observer who tries to guess whether there could be a car approaching from behind a curve before crossing a road: knowledge about that other car may currently be unavailable, but that car's state is correlated to her future experiences (namely, being hit by it or not).

Thus, the process that computes the measure μ can be interpreted as the observer's external world. But this does *not* imply that observers see actual bits, tapes, or other functional aspects of popular models of computation in their external world. The claim is only that there is an abstract probabilistic computational process that generates the observations, and our "external world" is a useful representation of this process. The mathematical framework on which our two postulates rely can be built on *any* kind of machine model. While in [1] I have chosen to work with a generalization of the monotone Turing machine, any other model that reads input bits and generates output histories sequentially will do equally well, and will define the exact same class of algorithms and probability measures and thus the exact same theory. For example, we could have built the framework on cellular automata, certain versions of λ-calculus, quantum Turing machines, or any other exotic machine model, as long as it can in principle be simulated (not necessarily efficiently) by a generalized monotone Turing machine and vice versa. Since the resulting theory is insensitive to the choice of model, the observer's emergent external world cannot be expected to resemble properties of any specific model of computation *except* for features that are common to all models of computation[12].

One such common feature is that computations must start in a simple initial state, and then complexity and entropy unfold in a simple algorithmic temporal evolution. This is exactly what we see in our world: extrapolating our universe's state to the past suggests that it has once started in a state of low complexity ("Big Bang").

There is yet another important feature of physics that we have to reconstruct: so far, every observer sees her private external world, and there is no apriori relation between the worlds of, say, observers A and B. For concreteness, suppose that observer A ("Abby") sees an external world in which events happen with probability ν (formally, ν is the probability distribution over the histories of the computational process as described above). For example, suppose that the probability that the sun is going to rise tomorrow in Abby's world is $\nu \approx 1$. But now suppose that Abby sees another observer, Bambi, in her world — some bunch of stuff that looks as if it encoded some mental states that describe the history of an observer according to our abstract definition. If Abby could somehow directly access Bambi's brain, she would find that Bambi has made a sequence of observations $\mathbf{x} = (x_1, \ldots, x_m)$ in Abby's world in the past. Abby could watch Bambi for a while, and see that Bambi makes some new observations $(z_1, \ldots, z_k)$. What is the probability that Bambi's next observation will be y? Well, according to Abby, it is $\nu(y|\mathbf{x}, z_1, \ldots, z_k)$. For example, the probability that *Abby will see Bambi seeing the sun rise tomorrow* will be close to one.

But according to our theory, there is another, apriori completely unrelated first-person probability as-

phenomenon. Thus, assuming that the convergence as in Theorem 2.2 is happening will hopefully give us insights that remain valid even if this assumption does not necessarily turn out to be satisfied exactly with unit probability.

[12] In more detail, these would be features that are common to all models of computation which are useful for observers that are part of the computation; see also the relevance of "predictivity" in Subsection 2.5.

sociated with these experiences: namely, algorithmic probability $\mathbf{P}(y|\mathbf{x}, z_1, \ldots, z_k; B)$. This is the actual chance of Bambi having next experience y. In principle, both probabilities have nothing to do with each other — it could be that Bambi will in fact see the sun rise tomorrow with probability much less than one ($\mathbf{P} \ll 1$), even though Abby sees Bambi seeing the sun rise with probability close to one ($\nu \approx 1$)! If this was the case, then Bambi would be a "probabilistic zombie" for Abby (adopting a vaguely related notion as introduced by Wittgenstein) — it would be an extremely counterintuitive situation in which Abby and Bambi would somehow not be "real" with respect to each other (take care that this is a quite colorful but highly problematic handwaving description). Good news: this does not typically happen.

Theorem 2.3 (Emergence of objective reality)**.** *The probabilities ν that determine the fate of Bambi (B) as seen by Abby are asymptotically close to the probabilities $\mathbf{P}$ which correspond to the actual propensities of Bambi's observations. That is, with ν-probability one,*

$$\nu(y|\mathbf{x}, z_1, \ldots, z_k) \overset{k\to\infty}{\longrightarrow} \mathbf{P}(y|\mathbf{x}, z_1, \ldots, z_k; B).$$

This is true under some assumptions [1, Assumption 10.1] which basically say that Bambi's experiences are well-defined in Abby's world for an arbitrarily long time in the future (i.e. Bambi is not terminated).

Thus, different observers will, after seeing each other for long enough, have compatible "chances of events" and in this sense become part of a common world. It is quite fascinating to see *why* Theorem 2.3 is true: it holds because of Solomonoff induction[13]. Namely, Abby's emergent external world, generating ν, is a computable probabilistic environment for Bambi; thus, if Bambi uses algorithmic probability $\mathbf{P}$ to predict future observations, then these probabilities will converge to the distribution ν according to SI. Theorem 2.3 relies on that exact same result, but reverses its interpretation. In other words, the same mathematical theorems that guarantee the correctness of induction algorithms in artificial intelligence [12] imply the emergence of objectivity in our theory[14].

2.4 Interlude: an open question

The two postulates on page 5 define a consistent theory that we will explore further below. Nonetheless, they have one major weakness that I do not currently know how to cure: namely, they claim that the probability of future observations y depends on the full past history $\mathbf{x}$ of the observer. It would be conceptually much cleaner and better motivated to postulate a *Markovian* probability distribution, i.e. a transition probability $\mathbf{P}(y|x)$ that only depends on the observer's momentary state $x = x_n$, *not* on the full past history $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. After all, in the absence of a fundamental notion of external world, the only significance of an observer's past should be in the memory that she holds about her past experiences, which will be encoded in $x = x_n$.

The main problem is that it is not clear how to define such a Markovian distribution (though there are a few natural candidates); for example, there is no notion of a "universal enumerable" Markovian semimeasure (for more of those information-theoretic details see [1, Open Problem 15.3 and below]). Moreover, this change of definition would make it impossible to import well-known results from algorithmic information theory (like Solomonoff induction), and so every attempt to prove interesting statements (like the ones of the previous subsection) would be much more difficult and mathematically challenging.

[13] In more detail, it holds because of a generalization of Solomonoff induction to the framework of observer graphs, histories, and graph machines, which can be found as Theorem 8.4 in [1].

[14] This is not the be confused with "Bayesian coherence", where observers (that are *by assumption part of the same world*) with different priors will have their posteriors converge towards each other after making a large number of compatible observations. The situation here is quite different: distinct observers have *actually different chances* of events to happen, and in this sense are not initially part of "the same world".

Open Problem. Is there a natural and simple alternative formulation of the theory in terms of a *Markovian* transition probability[15], i.e. a version of algorithmic probability $\mathbf{P}(y|x_1,\ldots,x_n;A)$ which depends only on the observer's current state x_n?

As long as we have not solved this open problem, and stick to the original formulation in terms of the two postulates on page 5, there will be a counterintuitive feature that can happen if there are *loops* in the observer graph. To see this, let us assume again that the event of Theorem 2.2 happens, i.e. that a simple computable measure μ determines our observer's chances asymptotically. The measure μ will not necessarily be Markovian (since $\mathbf{P}$ isn't), but we may still ask whether it allows an observer to predict her future observations $\mathbf{y}=(y_1,\ldots,y_m)$ if she only "knows" her current state $x=x_n$. (The problem is that the observer has no direct access to $x_1,\ldots,x_{n-1}$ and thus cannot determine $\mu(\mathbf{y}|\mathbf{x};A)$ directly.) To this end, let us call μ **predictive** if it uniquely induces a conditional probability distribution $\mu(\mathbf{y}|x;A)$ for all possible (momentary) observations x. Clearly, Markovian measures are predictive — by definition, $\mu(\mathbf{y}|x_n;A):=\mu(\mathbf{y}|x_1,\ldots,x_n;A)$ depends only on the final observation x_n. But non-Markovian measures can be predictive too, for example if they are *acyclic*, in the sense that the probability that the observer goes back to an earlier state is zero: $\mu(\mathbf{y},x,\mathbf{z},x;A)=0$ for all $x,\mathbf{y},\mathbf{z}$. In this case we have $\mu(\mathbf{y}|x;A)=\frac{\sum_{\mathbf{x}=(\ldots,x)}\mu(\mathbf{x},\mathbf{y};A)}{\sum_{\mathbf{x}=(\ldots,x)}\mu(\mathbf{x};A)}$, where the sums are over all histories $\mathbf{x}$ that end with x (convergence follows from acyclicity). However, if loops in the observer graph have non-zero probability, then the measure μ can be non-predictive, as the example of the **Sleeping Beauty Problem** in Figure 2 illustrates.

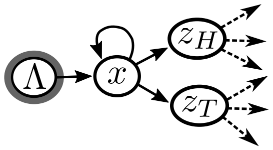

Figure 2: Sleeping Beauty Problem as described in [24]: *"Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads?"* This observer graph B describes Beauty's experiences: initially in some state Λ, at every waking in state x (which is attained once or twice), and then in a state where she is told that the coin toss was in fact Heads (z_H) resp. Tails (z_T). A measure μ on this graph describes the situation; it satisfies e.g. $\mu(\Lambda;B)=\mu(\Lambda,x;B)=1$ and $\mu(\Lambda,x,z_H;B)=\mu(\Lambda,x,x;B)=\mu(\Lambda,x,x,z_T;B)=\frac{1}{2}$. For Beauty, answering the puzzle is equivalent to coming up with a conditional probability $\mu(z_H|x;B)$, but no such probability can be uniquely inferred from the underlying probability space since, as it turns out, μ is not predictive. Beauty needs to come up with a purely subjective degree of belief on the "number of loopings" $x\to x$ to give an answer. See [1, Example 13.4] for details.

We can interpret predictive measures as those that admit an implementation of David Lewis' **"Principal Principle"** [25]: namely, that subjective degrees of belief should be determined by objective chance, if there is a notion of the latter. Once the open problem above can be solved, this will follow automatically in the amended theory, since if $\mathbf{P}$ is Markovian then μ is Markovian and hence predictive. These insights will become important in the next subsection.

2.5 What about quantum theory?

Conceptually, the theory as presented here is much more in line with quantum physics than the intuitions that we would get from a more traditional worldview (in fact, the phenomenology of quantum theory was a major motivation for its construction). Traditionally, we would think of a unique external world that evolves independently of any observer, which raises the question of why and how "measurements" could play any distinguished role whatsoever. Not so in the theory of this paper: there, it is in fact *observations*

[15] In standard algorithmic information theory, observers learn one new bit at every instance in time, which induces a Markovian measure. However, it is impossible to describe situations where an observer "forgets" information which I think is crucial.

which are primary, and the external world with its "elements of reality" which is emergent and secondary (in some sense, a very real-looking convenient fiction). This dissolves the measurement problem (similarly as in other epistemic interpretations of quantum theory, see [1, Observation 13.1] for a much more thorough discussion), and implements Peres' doctrine that "unperformed experiments have no results" [3].

While the theory of this paper does not directly predict the Hilbert space formalism of quantum theory (QT), it naturally leads to *some characteristics* of QT if we make the following additional assumption:

Assumption. For any given observer A, we assume that the measure μ which determines A's observations asymptotically according to Theorem 2.2 is *predictive, but not acyclic.*

See the previous subsection for a motivation[16]; in particular, if we can solve the Open Problem on page 11, then this assumption will automatically be satisfied by the amended theory[17].

So what does A's external world look like if it is determined by a simple computable predictive but not acyclic measure? Not being acyclic means that observers can run into "loops" like in the Sleeping Beauty Problem in Figure 2, which has some counterintuitive consequences. On the other hand, predictivity implies that the number of loopings $\#$ (which is either $\# = 1$ or $\# = 2$ in the Sleeping Beauty Problem, for example) can have no consequences for an observer's future experiences $\mathbf{y}$, given her current state x. This is because predictivity turns out to be equivalent to the conditional independence relation $\mathbf{y} \perp \#|x$. This implies that there are states of the observer's external world (i.e. of the computational process which generates μ) which are different (i.e. have distinct values of $\#$) but which are operationally indistinguishable for the observer — which can be seen as an instance of (preparation) **contextuality** as defined by Spekkens [26].

While this establishes a formal similarity with a characteristic feature of quantum theory, we should ask in more detail what this external world "looks like" for the observer that is contained in it. Before doing so, we have to clarify how we should really understand this question, and how we should not. It is a recurring theme of this work to emphasize that the question of "what is really going on in the world" is not a fruitful question to ask. Here this insight strikes again, since what we have is an (apriori abstract) computational process, i.e. an algorithm, and every algorithm can be represented in infinitely many ways.

To see why this is the case, and also to see how we can ask the question above in a meaningful way, consider an example computational world (or algorithm), namely *Conway's Game of Life*. It is defined as a two-dimensional cellular automaton, running on a square grid with squares that can be either black or white. The cells are updated in integer time steps according to simple rules (for more details, see an arbitrary popular online resource like Wikipedia or Gardner's text in Scientific American [27]), and the resulting behavior displays all kinds of complex and interesting behavior.

Yet, from an algorithmic point of view, there is nothing special about a two-dimensional cellular automaton: it is just one specific instance of a *model of computation.* There are many other (universal) models of computation like Turing machines, 1D cellular automata, Turing machines with several tapes, and so on, which are all algorithmically equivalent in the sense that any (universal) one of them can emulate any other. So if we regard Conway's Game of Life as an abstract algorithm (with bits that change their values over time), then this algorithm can be run on any one of those machine models — this is indeed what we are doing when we are simulating it on our desktop computers. Now, running Game of Life on a, say, (universal) 1D cellular automaton will be a quite peculiar thing to do. It can be done perfectly, but the simulation will be rather confusing to watch: after all, interactions of bits that are nearest-neighbor on the 2D grid have to be simulated by transporting information quite far and in a special way along the 1D automaton's cells.

While Game of Life can be represented (that is, run) on *any* machine model, the 2D cellular automaton model is one that is *particularly suited* for it: namely, the causal structure of the working of this machine

[16] As a further motivation, we can also say that if μ is not predictive then it says nothing useful for the observer; a "law of nature" that is useless or unavailable will simply not be recognized, period. On the other hand, one might also try to define something like the "predictive part" of a computable measure μ (basically all of μ that's useful for the observer to predict the future), and hope that the insights that we get by assuming full predictivity will remain generally true for the predictive part.

[17] In more detail, μ would be predictive since it would be Markovian. Generically, it would then also be cyclic unless it was finetuned to forbid loops.

model reflects in a particularly simple way the actual working of this algorithm. In other words, the 2D cellular automaton is a *natural model of computation* for Game of Life, and this insight expresses a characteristic feature of this algorithm.

Hence, the question we should ask in our context is the following: **What is the natural model of computation to represent algorithms that generate predictive (and possibly cyclic) measures μ?** The condition $\mathbf{y} \perp\!\!\!\perp \#|x$ mentioned above introduces severe constraints on the working of the model of computation. Suppose that the computer has a locality structure like a cellular automaton or a Turing machine (i.e. a notion of "closeness" of information — in a cellular automaton, for example, some cells are neighbors of others). Since the observer is part of the algorithm, some of her observations in the far future will depend on the values of other bits (seen as random variables) that are currently very "far away" in the computation. But fixing the values of these random variables early on will typically conflict with satisfaction of the predictivity constraint later. Thus, it seems that the computer has no choice but to *compute local processes for all different values of the random variables in parallel, and then later on to only "realize" one of the "branches" for the observer* in order to satisfy the predictivity constraint. In other words, it seems that we get a model of computation that shares some characteristic features with **Brassard and Raymond-Robichaud (BRR)'s (non-hidden variable) local realist model of quantum mechanics** [28, 29]: it will represent a sort of "many-worlds like" computational process, where only some of the branches interact or become visible for the observer — something remotely similar to quantum computation.

As shown in [28, 29], this sort of process will naturally lead to a **violation of Bell inequalities, but at the same time respect the no-signalling principle**[18]. Even without taking the BRR model into account, I show directly in [1, Section 13] that "non-locality but no-signalling" are generic properties that are seen by observers in predictive cyclic probabilistic worlds. The main idea is that "looping or not" (as in the Sleeping Beauty Problem, Figure 2) can depend on a random variable that is correlated with a distant random variable, which leads to an effective "detection loophole" in setups that resemble Bell scenarios. On the other hand, the predictivity constraint forbids observers to know whether they have looped or not, which generically[19] implies the no-signalling conditions.

This subsection represents the most speculative part of this paper. It relies not only on several mathematical theorems, but also on some plausibility arguments and e.g. the predictivity assumption. Thus, it should not be understood as the final claim about how quantum mechanics must come into play, but its main significance is to demonstrate that the theory of this paper is in principle *compatible* with quantum theory. It also gives a new perspective how to possibly understand some characteristic features of quantum theory. If there is indeed a grain of truth to the construction of this subsection, what can we learn about the interpretation of quantum theory? In a way, it would mean that all major interpretations of quantum theory are in some sense natural and valid viewpoints in this theory:

- On the one hand, the observer's emergent external world will correspond to a "many-worlds like" process[20] (though more in line with the BRR model than with an Everettian picture);
- on the other hand, there is actual "irreducible randomness" such as one sees in versions of the Copenhagen interpretation.
- But observers can still have lack of knowledge about their (non-fundamental) emergent external world,

[18] Note that there is a subtle loophole here: no-signalling is not really proven in generality in [28], and it is by no means trivial that it holds. Nevertheless, one can convince oneself pretty easily that the simple class of models that resembles [29] satisfies no-signalling due to "counting branches". Whether this is also true for the more complicated models arising in the context of this theory (and described in [1, Section 13]) is not automatically clear and thus needs separate analysis.

[19] For the exact assumptions under which this conclusion can be drawn, see for instance [1, Lemma 13.10]. On the one hand, predictivity implies that some random variables in the computation (like #) are forever inaccessible to the observer; on the other hand, we make an assumption that other random variables *do* generically become accessible to the observer in the future (at least with non-zero probability), which is what the word "generically" in this sentence refers to.

[20] Yet, note that the fundamental postulates on page 5 do not mention many worlds; they do not even mention a *single* world.

which will lead them to use nonclassical conditional probability distributions that resemble quantum states (or more general “nonlocal boxes” [1, Section 13]), in line with QBism.

Even if the theory of this paper should be completely wrong, I consider this a likely outcome of our future physical theories in general: namely, that all major interpretations of quantum theory will turn out to be valid at the same time in some sense, merely reflecting different viewpoints of human intuition on certain aspects of an unfamiliar ontology[21].

2.6 What about novel predictions?

So far, the discussion has focused on understanding known features of the world, but a good theory should also make novel predictions. By construction, these predictions will lie in the realm of problems mentioned in Subsection 2.1.

We have already seen in Subsection 2.3 how the theory resolves the **Boltzmann brain problem** in terms of a “principle of persistent regularities” (and we can learn more about “probabilities of possible worlds”, see [1, Observation 12.2]). Another quite surprising prediction of the theory is that objective reality breaks down in certain extreme situations. Recall Theorem 2.3 above: the notion of an “objective external world” is not a postulate of the theory, but rather a provable consequence — *under some assumptions.* Specifically, the fact that two observers A and B (Abby and Bambi) are “part of the same world” can only be guaranteed if both of them have made a large number of observations. If this turns out not to be the case, then a novel fascinating phenomenon can appear which is absent in all previous physical theories: namely, we will have a notion of **probabilistic zombies** (named after Wittgenstein’s zombies [30] which are a conceptually vaguely similar notion). This describes a situation where, for example, *Abby will see the sun rise tomorrow with probability close to 1, and Abby will also experience that Bambi will see the sun rise tomorrow with probability close to 1; however, Bambi’s chance of experiencing a rising sun will be much less than 1.* In this case, Bambi would be a probabilistic zombie. In fact, I show[22] in [1, Observation 11.1] that Bambi is a zombie for Abby if $K(\mathbf{x}) \ll K(\nu; B)$; where $K(\mathbf{x})$ is the complexity of the sequence of all of Bambi’s previous observations $\mathbf{x}$, and $K(\nu; B)$ is the complexity of the statistical distribution ν of Bambi’s observations which is induced by the “laws of nature” of Abby’s world. If Bambi makes more and more observations then $K(\mathbf{x})$ will grow indefinitely (and violate this inequality at some point since $K(\nu; B)$ is constant), but if Bambi is “very young” (or “very simple”) then she is a zombie (for Abby).

Another situation in which objective reality is predicted to break down is if Bambi is *terminated* in Abby’s world. In this case, the assumptions of Theorem 2.3 do not hold any more; in fact, the theory here predicts some sort of **subjective immortality** for all observers. For details of this, and how it relates to the ubiquity of *semimeasures* (instead of measures) in algorithmic information theory, see [1, Section 11].

The approach of this paper also makes statements about the problem of **brain emulation** as sketched in Subsection 2.1. The argumentation is laid out in detail in [1, Section 14], so I only give a very brief summary here. Suppose that we decide to simulate an observer to extreme detail on a computer — does the simulation correspond to an actual first-person perspective? This question also has an ethical dimension, as pointed out e.g. in [31]: for example, we might want to subject a simulated observer to potentially very painful medication tests in order to develop a cure for a terrible disease. Is this causing “real” suffering?

The theory as outlined above does in principle suggest a partial answer to this question, albeit one that may be hard to evaluate in practice. It tells us that the simulated sequence of observations $x_1, x_2, \ldots$ has an actual associated first-person perspective (an “observer”), and it allows us to determine how the

[21] Instead of asking “what is really going on in a quantum world?”, an arguably more fruitful question would be whether a general model of computation can be defined for predictive measures (as sketched above), and whether the resulting computational power of that model (and/or of the BRR model) would have any relation to quantum computation’s BQP.

[22] Note however that this argument in [1] is not yet formulated as a rigorous theorem but merely as a plausibility argument based on rigorously analyzed examples.

probabilities $\mathbf{P}$ of this observer's future experiences (for example, her probability of suffering) depend on how we construct and interact with the computer simulation. Similarly as the concept of a "probabilistic zombie" mentioned above, we might have (or aim for) a situation where the simulated observer has a high chance of suffering from an external point of view, but where this is not true for the actual chances in the corresponding first-person perspective.

In a nutshell, it turns out that the conclusion depends very much on the question *whether we have an "open" or a "closed" simulation*: if we simply program the computer and then let the simulation run but do not intervene in any way ("closed simulation"), then the act of running the simulation will have (almost) no consequences for any observer's actual first-person perspective. On the other hand, if we interact with the simulation in a substantial way (for example, by feeding information about the external world into it, or by communicating with the simulated observers) then this *will* have severe consequences for the associated observer. This conclusion follows basically from the difference in compressibility of the corresponding observations, which is the main property that is relevant for algorithmic probability. Thus, the approach of this paper could potentially be used to inform the ethical debate in this context.

2.7 What this theory is not

Since the theory is highly unconventional in the mix of areas for which it claims to make predictions, it is important to clarify what it is *not* meant to be:

> **Not a "theory of everything".** By construction, the theory suggested here will have to say nothing about most things. It will never be useful, for example, to say anything about quantum gravity or particle physics. We can see that the theory predicts its own limitations: according to Theorem 2.2 (and the way we have interpreted that theorem), a large part of what we call "laws of nature" will be contingent and impossible to predict by the theory in any detail. Instead, it is designed to approach questions in a well-defined realm, namely aspects of those that have been mentioned in Subsection 2.1.
>
> **No relation whatsoever to "consciousness" etc.** Even though there is talk about "observers" and "experiences", these are not supposed to correspond directly to any high-level aspects of our experience; in particular, no claim is made to say anything useful about notions like "consciousness". The binary strings which encode observations are to be understood in a technical way (relying on clear mathematical definitions), similarly as we can think of the "state of the brain of a bat" without any claims about "what it feels like to be a bat" [32]. In other words, there is no claim that this theory directly answers any questions related to the mind-body problem (though it is my hope that it can be indirectly useful for this field, as I alluded to in the previous subsection).
>
> **Not "yet another" interpretation of quantum mechanics.** Interpreting quantum mechanics is *not* a main goal of this work. Instead, the strategy was to learn from quantum mechanics as it currently stands, and to take its ontological implications seriously (in particular its implication that the notions of information, computation and probability have fundamental significance). The considerations of Subsection 2.5 above are only there to demonstrate that this work is compatible with quantum theory and may, under some assumptions, predict some of its characteristic properties.
>
> **No esoteric antirealism.** While I suggest that the physical world may be emergent instead of fundamental, I do not claim that this world is "not real", that everything is observer-relative, or that we should give up any of our standards for scientific theories. Quite on the contrary: the major goal here is to have a simple approach that allows to address questions rigorously that typically tend to be addressed with less rigor.

3 Conclusions

In this paper (and its more extensive companion paper [1]), I have sketched a simple and rigorous theory which predicts the emergence of an objective external world from a starting point that takes (only) an observer's first-person perspective as primary. As demonstrated above, the theory is well-motivated (for example, by the hypothetical success of Solomonoff induction) and has explanatory and predictive power.

There are several open problems and weaknesses of the theory as presented here. First and foremost, as elaborated in Subsection 2.4, it would be a major improvement to formulate the theory in terms of a *Markovian* version of algorithmic probability, but I do not currently know how to do this. Consequently, we had to make a "predictivity assumption" in Subsection 2.5 (discussing the relation to quantum theory), and some of the applications of this theory (for example in the context of brain emulation, see Subsection 2.6) will suffer from conceptual questions regarding the "current state" of an observer versus its "full past" (for more details see [1]). Furthermore, we have for the most part assumed that the event of Theorem 2.2 (convergence to some computable measure μ) happens for the observer, even though it is currently unclear whether this must happen with probability one. While one can argue that this probably captures important features that are also present if this event does not strictly happen (cf. the principle of persistent regularities), it would be beneficial to either prove convergence with unit probability or to analyze in more detail what non-convergence would imply. Along similar lines, it should be examined in more detail *how quickly* certain convergences take place, for example in Theorem 2.3 (emergence of objective reality), since such more finegrained results might provide ways to test and possibly falsify the theory. Finally, while the theory can be proven to be independent of the choice of universal reference machine (see [1, Subsection 5.3]), it is currently not clear how one would deal with that freedom of choice in practice and obtain actual numbers as probabilities of finite histories. This is arguably closely related to more general conceptual questions on the interpretation and test of probabilistic theories, which always seems to contain an irreducible element of subjectivity in some sense (cf. the arbitrary "five sigma" standard of particle physics).

Despite these drawbacks, we have arguably obtained a theory which has surprising explanatory and predictive power given its simplicity: the theory suggests partial answers to questions like *"why we do we see an external world that evolves according to simple probabilistic laws and seems to have started in a simple initial state?"*, or *"under what conditions does the computer simulation of an observer correspond to an actual first-person perspective?"*, even though the answer to the latter question will be more of a principled form and might be hard to evaluate in actual practice. While this theory may be completely wrong, it provides an undeniable **proof of principle** that we *can* construct a theory of this kind. As such, it points out that our approach to fundamental questions about the world may have so far suffered from a lack of imagination, and that our world could in a precise scientific sense be quite different from what we intuitively think that it should be.

If we accept this unlikely possibility for the moment, then we obtain a rather unfamiliar view of the world. Instead of a "world", we have observers that are on a never-ending journey through the mathematical space of possible experiences (i.e. their observer graphs). Sometimes observers will randomly find computable regularities in their observations, which then automatically stabilize themselves (due to the properties of algorithmic probability) and give observers the impression that they are part of an external world, governed by simple probabilistic "laws of nature". Observers meet, are part of the "same world" for a while, and then depart again. If we can solve the open problem on page 11, then the resulting Markovian process will either be transient (in which case observers will ultimately be led to more and more complex experiences), or recurrent. In the latter case, this could imply that there is in fact only a single observer: similarly as Marty McFly meeting his future self in "Back to the Future II", all other observers that we meet might simply be different instances of ourselves. I do not expect any of this to be literally true, but I hope that this work can open up fruitful new avenues of thinking about some fundamental questions.

Acknowledgments

I am indebted to many friends and colleagues for their comments, criticism and encouragement; please see the acknowledgments of the more comprehensive paper [1]. For this shorter version in particular, I would like to thank Philipp Höhn, Robert W. Spekkens, and in particular Michael Cuffaro for very helpful discussions. I am grateful to the Foundational Questions Institute (FQXi) for funding of the project "Emergent objective reality — from observers to physics via Solomonoff induction", and I would like to thank the Rotman Institute of Philosophy for funding via a Catalyst Grant. This research was supported in part by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Research, Innovation and Science.

References

[1] M. P. Müller, *Could the physical world be emergent instead of fundamental, and why should we ask? (full version)*, preprint (2017).

[2] M. Gardner, *The whys of a Philosophical Scrivener*, W. Morrow, New York, 1983.

[3] A. Peres, *Unperformed experiments have no results*, Am. J. Phys. **46**(7), 745–747 (1978).

[4] C. G. Timpson, *Quantum Information Theory & the Foundations of Quantum Mechanics*, Clarendon Press, Oxford, 2013.

[5] A. Albrecht, *Cosmic Inflation and the Arrow of Time*, in *Science and Ultimate Reality: From Quantum to Cosmos*, honoring John Wheeler's 90th birthday, J. D. Barrow, P. C. W. Davies, and C. L. Harper (eds.), Cambridge University Press, 2004.

[6] A. Albrecht and L. Sorbo, *Can the universe afford inflation?*, Phys. Rev. D **70**, 063528 (2004).

[7] D. N. Page, *Is our Universe likely to decay within 20 billion years?*, Phys. Rev. D **78**, 063535 (2008).

[8] A. Linde and M. Noorbala, *Measure problem for eternal and non-eternal inflation*, J. Cosmol. Astropart. Phys. **1009** (2010).

[9] D. J. Chalmers, *The Conscious Mind — in Search of a Fundamental Theory*, Oxford University Press, New York, 1996.

[10] N. Bostrom, *Are You Living In a Computer Simulation?*, Philosophical Quarterly **53**(211), 243–255 (2003).

[11] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, 2008.

[12] M. Hutter, *Universal Artificial Intelligence*, Springer Verlag, 2005.

[13] G. Piccinini, *Computationalism, The Church-Turing Thesis, and the Church-Turing Fallacy*, Synthese **154**(1), 97–120 (2007).

[14] J. A. Wheeler, *Law Without Law*, in "Quantum Theory and Measurement", ed. J. A. Wheeler and W. A. Zurek, Princeton Series in Physics, Princeton University Press, 1983.

[15] D. Parfit, *Reasons and persons*, Clarendon Press, Oxford, 1984.

[16] J. Leike and M. Hutter, *Bad Universal Priors and Notions of Optimality*, JMLR Workshop and Conference Proceedings **40**, 1–16 (2015).

[17] M. Müller, *Stationary algorithmic probability*, Theoretical Computer Science **411**, 113–130 (2010).

[18] J. Ladyman and D. Ross, *Every Thing Must Go*, Oxford University Press, 2007.

[19] C. M. Caves, C. A. Fuchs, and R. Schack, *Quantum probabilities as Bayesian probabilities*, Phys. Rev. A **65**, 022305 (2002).

[20] C. A. Fuchs, *Quantum Bayesianism at the Perimeter*, Physics in Canada **66**(2), 77–82 (2010).

[21] C. A. Fuchs, *Delirium Quantum Or, where I will take quantum mechanics if it will let me*, AIP Conference Proceedings **889**, 438 (2007).

[22] C. A. Fuchs, *Notwithstanding Bohr, the Reasons for QBism*, arXiv:1705.03483.

[23] A. Aguirre and M. Tegmark, *Born in an Infinite Universe: a Cosmological Interpretation of Quantum Mechanics*, Phys. Rev. D **84**, 105002 (2010).

[24] A. Elga, *Self-locating belief and the Sleeping Beauty problem*, Analysis **60**(2), 143–147 (2000).

[25] B. Weatherson, *David Lewis*, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL=`http://plato.stanford.edu/entries/david-lewis` (2014).

[26] R. W. Spekkens, *Contextuality for preparations, transformations, and unsharp measurements*, Phys. Rev. A **71**, 052108 (2005).

[27] M. Gardner, *Mathematical Games — The fantastic combinations of John Conway's new solitaire game "life"*, Scientific American **223**, 120–123 (1970).

[28] G. Brassard and P. Raymond-Robichaud, *Parallel lives: A local-realistic interpretation of "nonlocal" boxes*, arXiv:1709.10016.

[29] G. Brassard and P. Raymond-Robichaud, *Parallel lives: A local-realistic interpretation of "nonlocal" boxes*, poster (2015), available at `http://www.thepoxbox.com/tests/poster_revsmall.jpg`.

[30] R. Kirk, *Zombies*, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), ULR=`http://plato.stanford.edu/archives/win2012/entries/zombies` (2011).

[31] N. Bostrom, *Superintelligence*, Oxford University Press, Oxford, 2014.

[32] T. Nagel, *What Is It Like to Be a Bat?*, The Oxford Companion to Philosophy, 637, Oxford University Press, Oxford, 1974.