Emergent objective reality

From observers to physics via Solomonoff induction

Summary of a grant application, as a quick reading for participants of the "Participatory Realism" Workshop in Stellenbosch 2017. Please do not distribute further.

FQXi Large Grant Application: "Physics of the Observer"

Principal Investigator:

Markus P. Müller

Departments of Applied Mathematics and Philosophy, University of Western Ontario
Perimeter Institute for Theoretical Physics, Waterloo, Canada
markus@mpmueller.net

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 [1])

This proposal takes the perspective of the madman: it claims that this view of the world is wrong. Only if we doubt the obvious will we find answers to some important open problems in physics and beyond, including an answer to the question *why* we see something like an external world at all.

1.1 Postulates of an unusual but simple theory

Theoretical physics is more than just a collection of methods to predict measurable quantities like cross sections. Instead, the history of physics has given us plenty of examples when novel questions have led to new theories which fundamentally changed our picture of the world, often in surprising ways. An example is given by General Relativity, which has told us that our Newtonian view of an absolute and eternal space and time is only an approximation, and that in fact space and time can be curved and dynamical in a scientifically meaningful way.

The starting point of this proposal is the hypothesis that we are at a point where we perhaps have to make a comparably dramatic revision of some traditional aspects of our physical worldview. As I will argue below, there are several important conceptual problems in physics and its immediate vicinity that motivate such a move. *Traditionally*, a physical theory would start with an "ontic" picture of the world, postulating the existence of an objective external world that evolves according to certain physical laws. Our theories about these laws are tested by calculating some 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; 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 observe these probabilities as a consequence of the world's quantum state.

It turns out that several important conceptual problems in physics and beyond are closely related to the probabilities (1), and challenge this traditional view of physics:

- **Quantum mechanics** (QM). According to Bell's Theorem, naive versions of realism (roughly, the idea that measurement outcomes always exist 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" [2], and to decades of arguments about how to interpret the counterintuitive formalism of QM. Substantial interpretational effort has been invested in the question *"where the probabilities in (1) come from"*, without any final concensus.

- **Cosmology**. If we are observers in a really "big" universe (for example, a world undergoing eternal inflation), then the question arises 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 respect, for example the famous and notorious *Boltzmann brain problem* (claiming that we should assign high probability to us being only a short-lived quantum fluctuation), or, more broadly speaking, the *measure problem* of cosmology.

- **Artificial Intelligence / Philosophy of Mind**. 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 conceptual and ethical dilemmas. As one extreme and illustrative example, think of simulating the brain of a terminally ill person (after her death) on a computer. 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 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 project, 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 first sound, there is a clear-cut technical starting point, namely algorithmic information theory [3, 4] and "Solomonoff induction" (SI). In a nutshell, SI suggests that any observer who has made previous observations x should assign conditional algorithmic probability $\mathbf{P}_{\mathrm{alg}}(y|x)$ to future observations y. The quantity $\mathbf{P}_{\mathrm{alg}}$ is defined in algorithmic information theory as a "universal apriori probability". Roughly speaking, it corresponds to the probability that a randomly chosen computer program will output y if it has previously output x.

This prescription yields a method of inference that is provably optimal under many circumstances [4]. In particular, it is known to yield asymptotically correct predictions in *our* world, if we assume that the laws of physics are computable (as stated by the Church-Turing thesis). This is the reason why SI (or rather practically implementable versions of it) are actually being applied in Artificial Intelligence. In a nutshell, we can summarize the argument as follows (some more details will follow further below):

Definition of **Solomonoff induction:** Every agent that has made previous observations x should predict to make observations y with probability $\mathbf{P}_{\rm alg}(y|x)$ in the future.

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

In other words, we can in principle make 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 this case, we can instead try to use $\mathbf{P}_{\rm alg}$ for prediction as a pragmatic method of inference, motivated for example by considerations like Ockham's razor (since $\mathbf{P}_{\rm alg}(x)$ is larger for simpler x, i.e. for those that have a shorter description). This will in principle allow us to make predictions in realms where physics in itself does not, and is guaranteed to be compatible with physics in regimes where SI and physics are both applicable.

Second, it suggests a much more farreaching idea: *what if there is only one single "law of nature", 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 support a worldview that is completely different from the standard one, more similar to Wheeler's idea of "Law Without Law" [6].

Before examining this conceptual point in more detail, let us turn to the question how one can obtain a concrete theory from this idea. In this grant application, I will drop all mathematical details for obvious reasons; they can be found in [5], see the bibliography for a download link to a preliminary draft for the purpose of this FQXi grant application review only.

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 important features of it to allow for a mathematically sound theory. We will do this be introducing the notion of an "observer graph", which is a (computable, rooted) directed graph over the finite binary strings, $\{0,1\}^* = \{\lambda, 0, 1, 00, 01, 10, \ldots\}$. This captures the following idea. Any observer (say, a human being, or a robot) 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; maybe this string describes the experience and memory 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 would 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 possible next 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 "legitimate subjective 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 to materialize in the cave where the bat 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

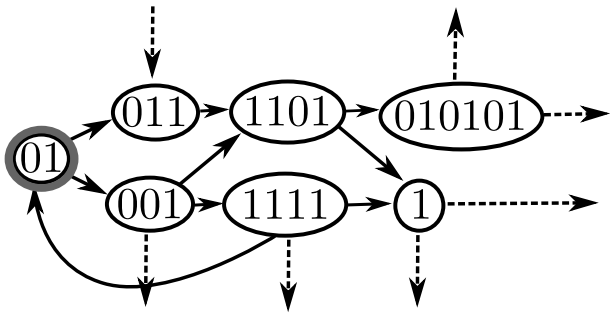

Figure 1: Schematic illustration of an observer graph A. Any path through the graph (starting at its root) will be called an "A-history". For any vertex x, we denote by $A(x)$ the set of all vertices y such that there is an arrow from x to y. For example, in this figure, we have $A(01) = \{011, 001\}$.

not be an arrow to a string that describes the experience of, say, Barack Obama at a state visit in Austria, looking up to the mountains (and having all his usual memories). This would not correspond to a legitimate subjective successor experience of the bat.[1]

Of course, the handwaving argumentation above raises all kinds of problems, 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 the theory will be "covariant" with respect to the choice of encoding, in a similar way as General Relativity is with respect to a choice of coordinate system. 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.

Postulates of the theory:

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

2. After having experienced a A-history $\mathbf{x} = (x_1, \ldots, x_n)$, the observer will subsequently experience one of the strings $y \in A(x)$ at random (cf. Figure 1). The probability of every $y \in A(x_n)$ is given by conditional algorithmic[2] probability $\mathbf{P}_{\text{alg}}(y|\mathbf{x})$.

This is it – there are no further postulates or assumptions. All other aspects of physics (including the appearance of an external world, and agreement between different observers) are not postulated, but rather 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.

[1] This concept is similar to Parfit's "Relation R" [9]. Nevertheless, the exact definition of this relation turns out to be irrelevant for the theory presented here.

[2] Conditional algorithmic probability $\mathbf{P}_{\text{alg}}(y|\mathbf{x})$ 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 sequences of strings as above (and this has been done in [5]). Technically, it is a generalization of the universal enumerable semimeasures on continuous sample spaces of [3].

Before we discuss how well-known aspects of physics follow from these postulates, we have to briefly discuss one problem that every reader with background in theoretical computer science will immediately acknowledge. By definition, $\mathbf{P}_{\rm alg}(y|\mathbf{x}) = \mathbf{P}_{\rm alg}(\mathbf{x}, y)/\mathbf{P}_{\rm alg}(\mathbf{x})$, and $\mathbf{P}_{\rm alg}(\mathbf{x})$ is defined as the probability that a universal computer outputs $\mathbf{x}$ if it is given a random program as input. *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 is regarded as an important problem in artificial intelligence [10] and does not have an easy solution [11]. However, it can be shown with some effort [5, Sec. 6.1] that the theory that follows from the postulates above is invariant with respect to the choice of U, as long as U is chosen from a specific infinite subset of universal computers.

1.2 Three roads to the same theory

Above, we have motivated our two postulates by the observation (framed on page 2) that Solomonoff induction's $\mathbf{P}_{\rm alg}(y|\mathbf{x})$ will give predictions which are asymptotically identical to those of our physical theories. From this, a kind of **"structural argument"** motivates our two postulates: if there is *one* canonical mathematical structure that gives us a notion of probability, or propensity, then we do not need another additional structure (another theory of physics) to explain these probabilities or propensities.

However, 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 **if there were no (predefined) laws of nature at all**. As I argue in [5], a way to formalize this idea is to say that "there should be no preference of any one choice of laws of nature over any other possible choice", and propensities of what happens (or rather what we see) should be determined simply by the structure of mathematics (containing all logically consistent possibilities) itself. Then one can try to come up with a "totally random choice of mathematical structure", and formalizing this idea will again lead to algorithmic probability, and finally to something similar to the two postulates above.

A third route is comparable to some ideas that cosmologists have come up with, cf. Aguirre and Tegmark [12]. Let us for a moment imagine the multitude of all conceivable, logically consistent worlds or "universes" (note that the theory above does not assume 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 [12]). 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 possible the moment before), then this should manifest itself as a random experiment for her. In other words, she should see a ***statistical mixture over all universes*** **consistent with her previous observations**. Now it can be shown without much effort [5] that algorithmic probability can be interpreted in exactly that way: as a statistial mixture over all *computable deterministic* worlds.[3] This gives us yet another way to interpret the two postulates.

1.3 So how do we get physics from this?

What would observers experience 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 $\mathbf{P}_{\rm alg}$. Even though $\mathbf{P}_{\rm alg}$ is a natural quantity, it is also very complex and irregular, in fact noncomputable. At first sight, this should imply that observers make completely irregular observations. However, I prove in [5] that this is not true: surprisingly, observers will see **simple, probabilistic "laws of nature"** in the long run.

[3] This is by no means inconsistent with quantum theory, as we will discuss further below.

That is, algorithmic probability $\mathbf{P}_{\text{alg}}$ converges to a random *computable* measure μ asymptotically, and the probability of a given μ is higher if its minimal description length (that is, Kolmogorov complexity K) is smaller. Let us look at the detailed mathematical theorem from [5], to have at least one single place in this grant application that gives a taste of the formal details (note that it is not necessary here to understand all details of the terminology):

Theorem 1.1. *Let A be a dead-end free observer graph, and μ a computable A-measure. Then*

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

that is, with probability at least $2^{-K(\mu)}$ (which is large if and only if μ is simple), the actual transition probability $\mathbf{P}_{\text{alg}}$ will in the long run converge to the computable measure μ (in Hellinger distance).

That is, after a while (i.e. after n observations, where n is large), observers will see that the randomness that governs their experiences converges to a "simple" probability distribution μ, i.e. $\mathbf{P}_{\text{alg}}(y|x_1,\ldots,x_n)$ (which is the actual propensity of a next observation y) will be very close to $\mu(y|x_1,\ldots,x_n)$, where μ is a "simple" probability measure. And "simple" means that there is a very short computer program, of length $K(\mu)$, that simulates this distribution. The shorter the description length of μ, the higher the probability to converge to μ. In other words, observers might say the following: *"It seems that what happens next is not completely deterministic... there is some randomness going on in the world. However, that randomness seems to be governed by simple probabilistic "laws of nature" (μ) that I can write down in very brief equations. Wow, that's interesting!"*

However, at this point, there is not yet any notion of "external world": exactly as $\mathbf{P}_{\text{alg}}$, the measure μ (the "laws of nature" that a given observer A experiences after many observations) is a probability distribution on the observer's states (that is, her knowledge and observations) only. However, I show in [5] that this distribution behaves *as if* the observer was part of a larger dynamical system which does not only consist of *her* state, but also of some other part that is not directly accessible to her. Taking this other part into account simplifies her task to predict future observations. Thus, every observer will find herself to be part of a computational universe, or **external world**, that evolves according to simple, computable, probabilistic laws. In a nutshell, this external world will simply correspond to *the computation*[4] *that generates the simple measure μ.*

Apriori, every observer will see her "own" universe; at this point, there is no notion of objectivity among several observers. To this end, consider the special case that an observer A ("Abby") observes another observer ("Bambi") in her external world – some bunch of stuff that *looks as if* it encoded some mental states that describe a possible other observer B. Apriori, the behavior of this part of Abby's external world is completely unrelated (according to the postulates of our theory) to Bambi's actual first-person experience: if Abby sees the sun rise tomorrow with probability close to one, then Bambi may in principle see the sun rise only with probability much less than one, *even though Abby sees Bambi experiencing a rising sun with probability close to one.*

However, I show in [5] that if Bambi is "old enough" (that is, if her state has large enough Kolmogorov complexity), then her subjectively perceived probabilities will match the propensities that Abby sees her experiencing in her own world. In other words, Abby and Bambi will asymptotically perceive to be "part of the same world", and agree on the probabilities of future events. This is a form of **emergent objective reality** (not to be confused with Bayesian coherence) which follows from the mathematics of Solomonoff induction. That is, the same mathematical theorems that guarantee the correctness of induction algorithms in artificial intelligence [4] imply the emergence of objectivity in our theory.

[4]Note that this does *not* mean that we should literally expect to see bits and bytes or memory tapes in our world, or even naive classical spacetime structure as in our standard desktop computers; it only shows that our universe should correspond to some abstract computational process; and, in fact, one that started out in a particularly simple state of "low entropy". This is exactly what we observe.

What about **quantum theory**? First, the conceptual picture painted by our theory seems much more compatible with quantum theory than our standard picture of a physical theory. Expressed in an exaggerated way, our theory says that *only observations* are fundamentally "real", and the external world with its elements of reality is merely a convenient fiction (but a very real-looking one, useful to predict future observations). This fits very well to the idea that "measurements" have a special status in producing actual outcomes, while "unperformed experiments have no results".

Yet, our theory has to say more about QM. As we have learned from quantum information and foundations research over the last few decades, any good foundational theory (that does not presuppose QM, but claims to predict it as a consequence of its postulates) must ultimately explain two facts:

1. The fact that we have a *violation of Bell inequalities* in our world, but at the same time the *no-signalling principle* holds [13] (i.e. no superluminal information transfer).

2. The fact that among *all* conceivable "probabilistic theories" [14] (including Popescu-Rohrlich boxes [15] etc.) that satisfy 1., we see exactly quantum theory, and not, for example, even stronger nonlocal correlations, higher-order interference [16, 17], or other conceivable effects.

Interestingly, the theory presented here explains 1., as shown in [5]: it predicts that whenever we have a positive probability of *loops* in the observer graph (as in Figure 1), the observer will see non-classical effects corresponding to "preparation contextuality" as defined in [18]. If the observer's external world admits a notion of locality, then this turns out to correspond to a violation of Bell inequalities despite no-signalling.

It is currently not clear whether our theory has anything to say about 2., and it is one of the goals of the research project proposed here to find out.

At this point, we have "reconstructed" several important aspects of physics as we know it. And in a way, the theory offers an explanation for "why" we see an objective external world at all. However, a theory should do more.

1.4 What about novel predictions?

A good theory should not only explain what we already know, but should also be able to make new predictions, possibly surprising ones.

In fact, the theory described above does so. The most surprising prospect is that it predicts **violations of objectivity** in some extreme situations. Remember from 1.2 above that the notion of an "objective external world" is not a postulate of the theory, but rather a provable consequence. But it only holds provably for two observers A and B (Abby and Bambi) if they are both "old enough", in the sense that their Kolmogorov complexity is high enough (this can be quantified in more detail, see [5]). If this is not so, then we will obtain a fascinating new phenomenon which is absent in all contemporary theories of physics – namely, we will have a notion of **probabilistic zombies** (different from, but named after Wittgenstein's zombies which are a conceptually somewhat similar notion). This would be 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 will subjectively only experience a rising sun with probability much less than 1.* In this case, Bambi would be a probabilistic zombie.

This phenomenon plays an important role in attempts to use this theory to resolve some of the open questions mentioned in 1.1 above. For example, it can rather easily **resolve the Boltzmann brain problem**, independently of any specific assumptions on the size or structure of our universe. While this is done in more detail in [5], let us give a quick-and-dirty overview on the main argument.

The argument starts with a colloquial interpretation of Theorem 1.1:

Persistance principle: Regularities that were holding in the past tend to persist in the future.

This is ultimately the reason why we have convergence to a simple measure μ in Theorem 1.1: if past observations happened to be compatible with μ, then algorithmic probability will favour future observations that are also compatible with μ.

With this principle in place, let us discuss how the Boltzmann brain paradox is automatically resolved. Suppose our observer Bambi is currently in a state where she remembers having lived a rich life full of experiences in a standard, low-entropic planet-like environment. However, 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 quantum 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 was very large (say, due to eternal inflation), then naive counting might seem to suggest that a BB-experience is in fact far more likely than our actual standard experience. However, according to our theory, Bambi's subjective experience is determined by algorithmic probability as in Postulate 2 (and *not* by counting frequencies), leading to the "Persistance principle" stated 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.* Thus, a BB-experience is extremely unlikely. In a nutshell, the reason is that a BB-experience has much higher Kolmogorov complexity than a standard experience.[5]

Returning to the "zombie" terminology from above, this would imply that actual Boltzmann brains, arising in quantum fluctuations (or other random processes), are in fact zombies in this sense. This can be analyzed in quite some quantitative detail [5].

This is just an example of novel predictions; others are related to the areas mentioned on page 2 (for example, brain emulation), and can be found in [5].

1.5 Isn't this completely crazy?

The theory presented here is definitely highly unconventional (and foundational, and topical since observers are the major starting point). Therefore, it makes sense to subject it to a quick sanity check. The following aspects make me confident that working on this theory is a fruitful endeavor:

No claim of a theory of everything. I am not at all claiming that this is supposed to be a "theory of everything" of some kind. There are definitely many aspects of physics that *cannot* be explained by this theory. This can already be seen in one of the first technical results, Theorem 1.1: according to our interpretation of this theorem, our laws of physics μ are at least to some extent a random coincidence. Apriori, all computable laws are possible, and all we can say is that there is a (strong) probabilistic preference for *simple* laws.

Concretely, this theory will never be useful in the search for a theory of quantum gravity, or in predicting cross sections of processes, etc. Instead, it is intended to admit possible answers to questions that are more fundamental than that, such as "why are there laws of nature at all?", or "what happens if we simulate an observer on a computer"?

Pragmatic interpretation. Instead of seeing Postulates 1 and 2 as the basis of a fundamental theory of our world, we can also see them as "pragmatic rules of thumb". This allows us to have a simple

[5] This needs some technical details to be spelled out quantitatively, cf. [5]. One has to be careful due to the difference between "the Kolmogorov complexity of a measure μ" (which is predicted to be small) and "the Kolmogorov complexity of a typical outcome of μ", which will typically be large. For example, think of μ as a billion coin tosses.

inductive method of reasoning in areas that are currently hard to understand from the perspective of physics (like the Boltzmann brain problem above), see also page 3.

Mathematical rigor. This is a mathematically fully rigorous theory. It is based on rigorous assumptions that are formally exploited to arrive at proven theorems. Some of the results are of independent interest in algorithmic information theory.

A most simple proof-of-principle. The theory is intended to represent a proof-of-principle that one can have a theory with explanatory and predictive power, which fundamentally starts with the notion of *observers* rather than of an external world. No matter whether one believes that it is true or not, it is valuable to see that this is in principle possible. Since the two postulates (cf. page 4) are so simple, this makes it particularly transparent and an important starting point for further research that shares some of its ideas with our theory.

Fruitful spin-offs. This theory gives some fruitful ideas that can be pursued detached from the theory itself. For example, it suggests to use algorithmic probability as a measure in cosmology, and it indicates that a notion of "fundamental forgetting" (loops in observer graphs) can be responsible for the observation of quantum nonlocality and no-signalling, see also Section 2.

In any case, the ideas presented above give motivation to think about our world in novel terms, and show that a scientifically rigorous theory can have a surprisingly different form from what we are used to.

2 Acknowledgment

I am deeply grateful to my colleague and FQXi postdoc Michael Cuffaro, who is now working with me on this project (a paper on the relation to Carnap's "Aufbau" is in preparation). I have been (and am) learning an invaluable amount of philosophy of physics from Mike, and this project has benefitted immensely from his expertise.

3 Bibliography

References

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

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

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

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

[5] M. P. Müller, in preparation.

[6] 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.

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

[8] C. Fuchs, *Quantum mechanics as quantum information (and only a little more)*, in A. Khrenikov (ed.) Quantum Theory: Reconstruction of Foundations (Växjö: Växjö University Press, 2002)

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

[10] J. Leike and M. Hutter, *Bad Universal Priors and Notions of Optimality*, arXiv:1510.04931.

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

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

[13] C. J. Wood and R. W. Spekkens, *The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning*, New J. Phys. **17**, 033002 (2015).

[14] J. Barrett, *Information processing in generalized probabilistic theories*, Phys. Rev. A **75**, 032304 (2007).

[15] S. Popescu, *Nonlocality beyond quantum mechanics*, Nat. Phys. **10**, 264–270 (2014).

[16] R. D. Sorkin, *Quantum mechanics as quantum measure theory*, Mod. Phys. Lett. A **9**, 3119–3127 (1994).

[17] U. Sinha, C. Couteau, T. Jennewein, R. Laflamme, and G. Weihs, *Ruling Out Multi-Order Interference in Quantum Mechanics*, Science **329**, 418 (2010).

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

[19] J. Eisert, M. P. Müller, and C. Gogolin, *Quantum measurement occurrence is undecidable*, Phys. Rev. Lett. **108**, 260501 (2012).

[20] Ll. Masanes and M. P. Müller, *A derivation of quantum theory from physical requirements*, New J. Phys. **13**, 063001 (2011).

[21] H. Barnum, M. P. Müller, and C. Ududec, *Higher-order interference and single-system postulates characterizing quantum theory*, New J. Phys. **16**, 123029 (2014).

[22] C. A. Fuchs and R. Schack, *Quantum-Bayesian coherence*, Rev. Mod. Phys. **85**, 1693–1715 (2013).

[23] J. Ladyman and D. Ross, *Every Thing Must Go*, Oxford University Press, Oxford, 2010.

[24] P. M. Ainsworth, *What is ontic structural realism?*, Studies in History and Philosophy of Modern Physics **41**, 50–57 (2010).

[25] A. Zuboff, *One Self: The Logic of Experience*, Inquiry: An Interdisciplinary Journal of Philosophy **33**(1), 39–68 (1990).

[26] J. Schmidhuber, *Algorithmic Theories of Everything*, Instituto Dalle Molle Di Studi Sull Intelligenza Artificiale (2000).

[27] S. Lloyd, *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*, Random House, New York, 2006.

[28] M. Tegmark, *Does the universe in fact contain almost no information?*, Found. Phys. Lett. **9**, 25–42 (1996).

[29] S. Lloyd and O. Dreyer, *The universal path integral*, Quant. Inf. Proc. **2**, 959–967 (2015).

[30] Y. Benétreau-Dupin, *Probabilistic Reasoning in Cosmology*, PhD Thesis, University of Western Ontario, 2015.