Consciousness in Wolfram Models

by Sven Nilsen, 2021

In this paper I outline an idea of how consciousness might work relative to Wolfram models.

A Wolfram model[1] is a hypergraph[2] rewriting system which assigns every natural number a unique rule and an initial condition, with labeled identical nodes and hyperedges, such that the hypergraph evolution can be studied without knowing anything more than the associated number.

I consider the most important insight about the universe from a perspective of Wolfram models to be the ability to define physical measurements (SI units[3]) over the language of hypergraphs. These definitions work universally and can be tested over a large class of models.

One problem of SI units, is that scientists do not know yet how to bridge the explanatory gap between mental states[4] (e.g. the experience of consciousness) and physical systems. The Integrated Information Theory of Consciousness[5] suggests that mental states always occur in presence of information processing, but this theory is only measuring density of consciousness as a scalar value.

To derive how mental states might occur in Wolfram models, I will start with the language that is used in the theory of Avatar Extensions[6] and gradually bridge the gap between the two.

$a \cdot b$	An abstract product of `a` and `b` in Avatar Algebra[7]

Let `a` be some form of consciousness and `b` some kind of information processing. The product represents the experience of awareness. Now, I will make the assumption:

1. The potential for awareness exists originally prior to any physical interaction

For example, the potential to experience a red color, exists prior to the experiencing of seeing a red color. While this might seem like just a game design of language, it is important because how one thinks about awareness relative to physical systems. Since this potential exists prior to the realisation of awareness, one can simplify the representation of the potential to discrete states.

Simplified:

```
seeing_red
seeing_blue
…
```

Physical systems are like edges between nodes of awareness. Since this introduces a cyclic language problem where experiencing red in one context can not be distinguished from seeing red in another context, I will use 1-avatars to relax the language constraints:

seeing_red_0	seeing_red_1	seeing_red_2	…
seeing_blue_0	seeing_blue_1	seeing_blue_2	…
…	…	…	

A 1-avatar “wraps” the core, which in this case is `seeing_red` and `seeing_blue`. The number of wrappings correspond naturally to the Peano arithmetic of natural numbers[8]. Such 1-avatars can be constructed systematically in a Turing complete system. This mean that, although the potential for

seeing red exists prior to the realisation of seeing red, there is another view in which each experience of seeing red is unique. This view is “outside” the 1-avatar. The experience of seeing red differs only from other experiences of seeing red by the context which this experience takes place. The context is represented as the wrapping of the 1-avatar, which is a systematic construction.

The total physical laws of the universe are unknown. Yet, one can reason about them from the perspective of a language which talks about physical laws in some set of possible universes. When the wrapping of 1-avatars happens, it is not necessarily that this should be interpreted literally as what is going on in the physical laws. The idea of 1-avatars is a language tool to assist the reasoning about systematic constructions. Basically, the idea when there is only one way to create something new out of something that already exists. The realisation of the potential of awareness into a concrete experience is the systematic construction by physical laws.

The potential for awareness exists prior to anything happening in the universe. This does not mean that the potential has the form of space-time where everything is static. Instead, the potential for awareness might be thought of as an “original object” which is immutable[9] and causes awareness when it is constructed “inside” a 1-avatar.

Here, I made a second assumption:

> 2. The awareness in a context of physical systems is uniform of the original potential

With other words, the potential for awareness in space and time does not change with its amount of physical interactions in the context. For example, there is no universe in which the potential for awareness goes to zero while the same structure of physical intractions are possible. The potential for awareness is the same, always and everywhere. However, this potential must be seen as separated from the potential of physical interactions. Only the combination of the two gives the potential of realised experience. Since the potential of awareness is the same everywhere, the potential of realised experience depends only on the potential of physical interactions.

Without the 1-avatars of the “original” potential for awareness, the language of physics would be stuck in infinite loops leading back to the same initial condition. This would be like seeing red always and everywhere, while simultaneously seeing blue always and everywhere and no ability to distinguish between the two, since there is no physical context to separate them.

To simplify this model further, I would like to neutralise the potential for awareness. This is done by zipping all the possible 1-avatars into a single sequence of natural numbers:

0	**1**	**2**	**3**	**…**
seeing_red_0	seeing_blue_0	seeing_red_1	seeing_blue_1	…

Notice the similarity to even and odd natural numbers. The point is to put things together which are by their very nature different from each other. This allows one to ignore how each discrete state is systematically constructed. This process removes bits of information that distinguishes potential of awareness. However, given access to some algorithm to restore this information, one can identify the 1-avatar in the old sequence.

Now, I make a third assumption:

> 3. There exists a Wolfram model such that the neutralised sequence of 1-avatars corresponds to the nodes in the hypergraph evolution, where the physical context of the node contains the information necessary to reproduce systematic construction of 1-avatars

It is not the physical interactions that “cause” awareness directly, but the systematic construction of 1-avatars. Yet physical interactions determine the form of awareness. The nodes are always there to be assigned some additional meaning. This means that the action that assigns something to nodes is always present. In many cases, one can treat the physical interactions as same as the form of realised awareness. Physical interactions are like pieces of information used to assign something to the node. One consequence of this view is that a simulated world does not necessarily give arise to same mental states as the physical world. The algorithm that assigns something to the node might do a different action depending on the physical context, which differs in simulations.

The action that assigns something to the node identifies the 1-avatar of that particular realised awareness. I make this design choice to help intuition about how awareness occurs. This is seen as a separate language mechanism than the physical interaction itself. The reason for this is that when humans talk about awareness, there is no language to distinguish mind states in full detail. The language used to talk about awareness and consciousness is limited to inaccuracy.

Neutralised 1-avatar Restored 1-avatar

The process of assigning something to nodes “causes” awareness

If this is sufficient for talking about experiences of awareness and doing experiments, then there is no reason to make the language model more complex. Avatar Extensions is helpful to explain how the language about consciousness works and how to integrate this language with Wolfram models.

Even the potential for awareness exists originally prior to any physical interaction, this does not imply that the potential for awareness is finite. The original potential might very well be like a fractal[10] (suggested by Daniel Fischer).

The information from the physical context might be used to look up the systematic construction of the 1-avatar in a way that is undecidable for small changes in the parameters. However, for other states of awareness, even large changes might not affect the kind of awareness very much.

A seemingly problem is that since awareness is intrinsically subjective, it can be difficult to translate from the experience of one person to another. However, the information from physical context can be used to determine the properties of the kind of awareness. Without this relation, it would be impossible to even talk about consciousness in the first place.

One critical view of assignment of something to nodes in a Wolfram model is Occam’s razor[11]. Why should I suggest a language with additional entities beyond those that are physically measurable in SI units?

The reason is that I assume SI units can not account for all aspects of existence, such as states of mind, so there must be some extra information in Wolfram models to distinguish between states. The information is not extrinsic to the hypergraph, but derived from it and asserting the physical existence of these properties from a subjective perspective. Objectively, the derived subjective experience is defined, but only experienced subjectively. However, it needs to be “stored” somewhere. This information can either be assigned to hyperedges or nodes. If one chooses hyperedges, then this information can be projected onto information about nodes. Likewise, if one chooses nodes, then this information can be projected onto information about hyperedges. The reason I choose nodes is that it is conceptually simpler to reason about compared to hyperedges.

It is always possible to assign this information to nodes, since the state of the evolving hypergraph is well defined.

Furthermore, it is important to not over-simplify this idea. No matter how many unknowns there are in the theory of consciousness, there are some obvious aspects such as consciousness not being associated with a simple physical state. The “seeing red” and “seeing blue” are kind of toy examples which can not be narrowed down to a single node in a Wolfram model easily.

On the other hand, I believe that in order to talk properly about agents experiencing their environment, as in how internal states of mind are generated (as opposed to the accuracy we use in natural language to talk about consciousness), the minimum language required is quantum circuits. This might require Homotopy Physics[12].

In the case of individual nodes in a Wolfram model, there are two metaphysical interpretations:

1. Each node being some kind of “primitive” realisation of awareness
2. Each node being an abstract realisation of awareness

The second kind of metaphysical interpretation can be thought of as a normal path:

    avatar_model[neutralisation_model] <=> core_model

The core model is a simplified physical model of the universe. Each node corresponds to some unit of realised awareness. A neutralisation model is used to contract the actual avatar model down into the core model. The choice of units of awareness is determined by the neutralisation model.

Hence, it is possible, even if a Wolfram model is discovered that accounts for all physical phenomena measurable in SI units, that it is unable to account for consciousness and awareness. This is because although the information is derived from the same hypergraph, the correct choice to do so is unknown.

Since the abstract view is possible, it might also happen that consciousness in Wolfram models must be viewed as an emergent phenomena which can not be reduced without missing something. This is because the Wolfram model is some sort of closed language which can not actually refer to reality, but requires an outside observer to be used in the correct context.

References:

[1] "A Class of Models with the Potential to Represent Fundamental Physics"
Stephen Wolfram
https://arxiv.org/ftp/arxiv/papers/2004/2004.08210.pdf

[2] "Hypergraph"
Wikipedia
https://en.wikipedia.org/wiki/Hypergraph

[3] "International System of Units"
Wikipedia
https://en.wikipedia.org/wiki/International_System_of_Units

[4] "Mind-body problem"
Wikipedia
https://en.wikipedia.org/wiki/Mind%E2%80%93body_problem

[5] "Integrated Information Theory of Consciousness"
Internet Encyclopedia of Philosophy
https://iep.utm.edu/int-info/

[6] "Avatar Extensions"
AdvancedResearch – Summary page on Avatar Extensions
https://advancedresearch.github.io/avatar-extensions/summary.html

[7] "Avatar Algebra"
AdvancedResearch – Summary page on Avatar Extensions
https://advancedresearch.github.io/avatar-extensions/summary.html#avatar-algebra

[8] "Peano axioms"
Wikipedia
https://en.wikipedia.org/wiki/Peano_axioms

[9] "Immutable object"
Wikipedia
https://en.wikipedia.org/wiki/Immutable_object

[10] "Fractal"
Wikipedia
https://en.wikipedia.org/wiki/Fractal

[11] "Occam's razor"
Wikipedia
https://en.wikipedia.org/wiki/Occam%27s_razor

[12] "Homotopy Physics"
Advanced Research – Reading sequences on Path Semantics
https://github.com/advancedresearch/path_semantics/blob/master/sequences.md#homotopy-physics