I think, therefore I think you think I am

Sofia Magnúsdóttir

Conceptual clarity is the foundation of scientific discourse. Therefore, I wish to propose a new way to speak about and quantify consciousness. This new definition is based on the ability of a system to accurately monitor and predict its environment and itself. While I am at it, I will also explain philosophical zombies, free will, and the purpose of life.

The Big, Bad Question

"What is consciousness?" isn't only a big question, it's also a bad question. That's because any consistent definition would answer it. Here is one: "Consciousness is a lemon tree." That, you might complain, is not what you mean by consciousness. Too bad, then, that you cannot tell me what you mean because you don't know what consciousness is. What a mess.

The Goal

To make headway on this big, bad question, I therefore first have to make the question more concrete. I am looking for a definition that captures what most of us (humans, presumably) mean by consciousness, a definition according to which a) rocks aren't conscious but b) most animals are, c) animals can be conscious to varying degrees, and d) even be temporarily unconscious. Besides fulfilling the requirements a)-d), the definition should also enable us to answer following three representative questions:

Q1) Is an anesthetized person safely out so that they do not experience pain?

Q2) Is a person with locked-in syndrome self-aware and/or aware of their situation?

Q3) Has an artificial intelligence developed consciousness comparable to that of animals?

Finding a definition that fulfils requirements a)-d) and allows answering questions Q1-Q3 is surprisingly difficult. The most common way to test for consciousness is to seek a reaction, for example by prodding a patient or checking pupil contractions. But this probes for more than just consciousness because it moreover requires visible output.

We might want to assert that certain reactions are, if not necessary, then at least sufficient indicators for consciousness. This assertion is questioned by the concept of a "philosophical zombie," an imaginary being that reacts like a human yet does not have experience[1]. But is it even possible for someone (some thing?) to behave like a human without also having human-

like experiences? This is another question which a good definition of consciousness should be able to settle.

A somewhat more advanced test for consciousness might scan brain activity. But this does not help us much either because we don't know which type of brain activity demonstrates experience or consciousness. Again, we might want to search for a reaction to stimulus, but any computer with an input/output (hereafter I/O) interface, such as a camera or simply a keyboard, does – in some sense – 'react' to input. Most of us, however, do not ascribe consciousness to present-day computers.

Previously proposed output-independent measures for consciousness have focused on a system's capability for and type of information processing, like Tononi's Integrated Information Theory[2] or Tegmark's Consiousness as a State of Matter[3]. These approaches suffer from two problems. First, it isn't clear whether they actually capture what we mean by consciousness, but – in all fairness – these are early days for consciousness models and settling the issue might just take more time. More importantly, however, approaches that aim to quantify consciousness based on a system's structure leave us wondering what the system is supposedly conscious of.

So, yes, it's a hard problem, but if it was easy then what would be the point of this contest? The goal of this essay is hence to answer the above questions. Since I am not a neurologist, however, the reader be warned that I only offer an answer 'in principle,' not an answer 'in practice.' I believe, however, that with further refinement the approach I want to propose can become practically useful (see Appendix).

Clues from Evolution

Let us begin by asking how we even got to the point of asking "What is consciousness?" We are products of Darwinian evolution. 'Survival of the fittest' is commonly interpreted as an adaptive selection of actions beneficial for reproduction. But this pays too little attention to the question what it takes to develop these reactions.

Take, for example, a rabbit. A rabbit which runs or hides when it sees a tiger has an evolutionary advantage over a rabbit that mistakes the tiger for a carrot – that much is clear. Less clear is what the rabbit brain must do to recognize the threat. (I'm not sure there are many rabbits in the natural habitat of tigers, but you get the point.)

To begin with, the rabbit brain must be able to obtain sufficiently accurate information about its environment, which means it needs an input channel. More importantly for our purposes, it must be able to identify patterns matching the threat 'tiger.' This means the rabbit brain must be able to create a reasonably accurate model of its environment: It must have an internal

representation that faithfully encodes information about the environment, and further a way to process this representation to arrive at a reaction.

This internal representation allows for reliable if-then reactions, which is a good starting point. But the rabbit can become even 'fitter' when it uses more sophisticated models, which take into account not only the present state of the tiger but extrapolate the tiger's behavior, possibly including also the reaction of the environment. Does the tiger look like it has seen the rabbit and is about to jump? If it jumps, where will it likely land? What's the tiger-ground friction coefficient? If the rabbit has a predictive model that computes faster than the environmental story unfolds, it will allow the rabbit to get out of the way. Accurate predictions, therefore, are plausibly a survival advantage and hence a likely product of evolution.

Indeed, there is much evidence that the brain is good at exactly these tasks. Our ability to extrapolate trajectories seems at least partly genetic. Mouse brains are known to have 'place cells' corresponding to their location in a maze[4] and various species have been found to have brain structures encoding small integers[5]. All these are examples of environmental models. Consequently, pattern recognition is what neural networks are now trained to perform at, in mimicry of human development.

Terminology

Let us be precise with the terminology by employing the language of mathematics. One may debate whether consciousness can be entirely captured by mathematics, but this need not worry us here. We merely note that mathematics has been successfully applied to categorize and understand many natural systems, and it therefore seems reasonable to also use it to better classify consciousness.

By **system** we will simply mean a set with distinct elements that have properties and relations among each other. A **subsystem** is a subset of the system. We'll refer to the largest possible system as **universe**. If the system under consideration is not identical to the universe, we call its complement **environment**. Remaining connections between a system and the environment represent input/output channels.

Note that in these definitions we have not made any assumptions about the spatial relations between the elements or their compactness. We will come back to this later.

The reductionists among the readers may want to think of the system's elements as elementary particles and the relations as interactions, but there is no need for this particular interpretation. The system's elements can as well be emergent objects in a higher-level theory. However, I wish to emphasize that it is not necessary to identify the elements of the system with synapses connected by axons, they could as well be circuits or qubits.

We will use the word **model** to mean a morphism, that is a map between elements which preserves the relation between the elements and the assignment of properties. Ie, two elements that have the same property (relation) P in the origin have the same property (relation) P' in the image, though two elements that had different properties (relations) P and Q might end up having the same property (relation) P'. In other words, the morphism is not, in general, an isomorphism.

Since nothing real is ever perfect, the morphisms we will deal with here generally have a limited fidelity. By this we mean that the image of the morphism doesn't preserve the relations or properties of all elements correctly. We will hence speak of **imperfect morphisms**.

The fidelity of the morphism quantifies the model's accuracy. There are various ways to quantify fidelity – we could for example just count the number of mistakes in the properties and elements, and divide them by the number of properties and elements. There is no unique definition for fidelity but for our purpose we only need to know that it is a property which *can* be quantified. That such a quantification isn't unique means that the absolute value of a model's accuracy isn't meaningful, but we can still make relative comparisons. (Similar to the case with the utility function employed in equilibrium economics.)

The system and the environment will further in general be dynamical, meaning they will change over time. If the morphism is also time-dependent but achieves to model a time-sequence of states of the environment faster than the change in the environment occurs, we will refer to it as a **predictive morphism**. Again we could quantify how good the model's prediction is by measuring the mistakes that occur over some time[6].

Four Levels Of Awareness

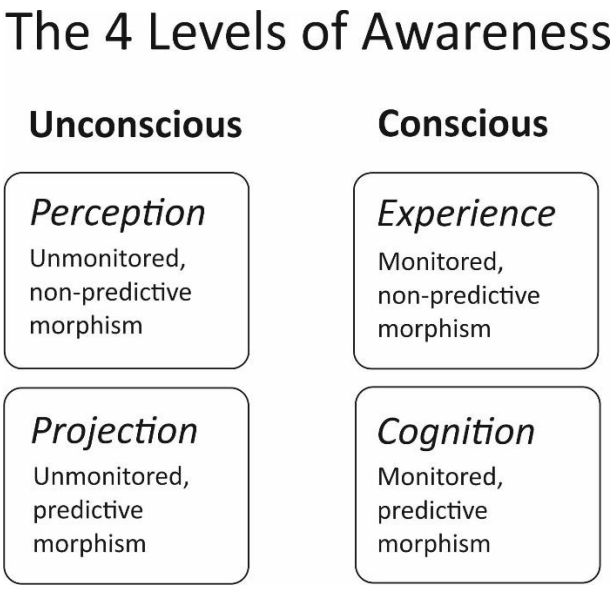
Armed with this terminology, let us return to the rabbit. We have seen that a system has better chances of survival when it has an accurate model of the environment, better still if the model is moreover predictive. The rabbit only benefits from this if it is able to do anything in reaction, like run away, but the model is a prerequisite for this.

The prediction of unfolding events, however, will be even better if the rabbit takes into account also its own actions and their likely consequences. Therefore, the rabbit increase its evolutionary advantage if it understands its likely reaction to a certain input. In other words, the rabbit will perform better if its brain has a model of itself modeling.

This brain self-model can, again, either merely monitor the rabbit's modeling – a non-predictive morphism – or it may be predictive. The former case we will refer to as 'experience,' the latter

as 'cognition.' Finally arriving at the concept of consciousness, we will refer to the two cases which include a self-model as 'conscious awareness,' whereas the predictive and non-predictive models without self-monitoring represent 'unconscious awareness.'

This leads us to the following four-level classification of consciousness (see Figure left):

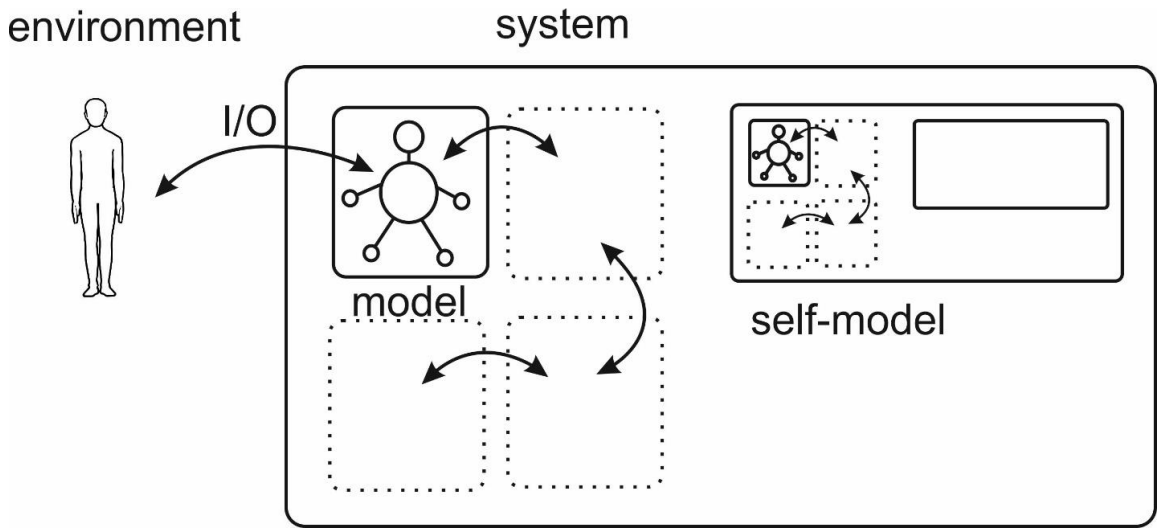


1. **Reception:**
The system receives input and produces a model of its environment. This model will in general have a limited fidelity. The better the fidelity the higher the awareness of the environment.

2. **Projection.**
The system has a predictive model of its environment. Predictive models are only useful if the prediction of an event can be made before the event happens.

3. **Experience**
The system contains a subsystem with a self-similar model of itself that monitors the integration of input and its internal connectivity possibly including the monitor itself. This is akin the task-manager of an operating system.

4. **Cognition**
A system has cognition if it contains a subsystem with a self-similar predictive model of itself, again possibly including the monitor itself. (See Figure below.) The self-model, importantly, not merely again images the environmental model, but must also model the connections between the subsystems that models (a certain part of) the environment and other subsystems (dotted).



The ability of the human brain to model its own self-modeling is presently poor. This is the very reason we're having this essay contest.

Learning from That

An immediate consequence of the above proposed definition is that a system isn't either conscious or not, but it can be conscious to varying degrees. Consciousness, hence isn't binary but continuous. How conscious a subsystem is of its environment or other subsystems, depends on how good a model it has, how predictive the model is and how well it is monitored. The fidelity of the morphisms therefore quantify conscious and unconscious awareness.

A more important lesson, however, is that consciousness is not in and by itself a property of a system. Instead, consciousness is relational: Its origin lies in the relation between a system, its environment, and its subsystems.

Consciousness, hence, is a noun that is shorthand for a verb much like, for example, the word "leadership". Leadership too isn't a thing and it isn't a property, it's a relation. You can't just lead, you can only lead somebody or something. Consciousness, likewise, isn't a thing, it's a relation. You're not just conscious, you are always conscious *of* somebody or something.

It is well possible that the structurally-based models of consciousness like the ones proposed by Tononi and Tegmark identify exactly the kind of systems which can achieve exactly the model-building discussed above here. I believe, however, that focusing on self-similar maps is a more minimalistic way to think of consciousness.

A further lesson from the above is that consciousness is extendable: Since consciousness is the ability of a system to comprehend another system, consciousness can be expanded by supporting this comprehension externally. Moreover, as anticipated above, there's no particular reason why a conscious system needs to by compact or consist of components close by each other to create a predictive model. The components should be connected so as to exchange information, but other than that consciousness could extend over long distances. However, we may suspect that information-processing in very extended systems will be too slow and inefficient so that the evolution of consciousness is unlikely.

A point which we have not addressed above is the irreducibility of the conscious system. According to the above, any system that contains a conscious subsystem is also conscious, which is clearly not the way that we think about consciousness. Therefore, we should complete our definition by requiring that removing any elements or connections from the system will significantly decrease the system's level of consciousness (fidelity of the model).

To be more concrete on this point we would first have to distinguish connections between elements which carry information from those which carry supply (say, blood). At present,

however, it isn't possible to clarify this definition because nobody knows how much of cognition may be 'embodied.' Indeed the definition proposed here could aid in identifying embodied cognition.

In summary, we have learned that consciousness is continuous, relative, extendable, and distributable.

Self-Check

Let us now see whether we met the goal we set for ourselves.

a) Rocks rarely change internal states, hence cannot create models of their environment, at least not in the typical lifetime of solar systems. A simple self-similar system can, to some extent, be said to have an experience of itself, but so long as it's not an experience of anything in particular, we probably wouldn't ascribe it much meaning. Having said that, it's interesting to note that self-similarity and scale-invariance are indeed hallmarks of complex systems and a relation to consciousness has been suggested previously[7].

b) Animals can, to varying degree, model their environment as evidenced by their ability to react and can therefore safely be said to be aware of their environment. The typical test for self-awareness – that an animal is able to recognize its reflection in a mirror – also beautifully fits into our definition. Recognizing one's reflection means that one has been able to identify part of the environment with internal processes, which requires self-monitoring and integration of one's own environmental modeling with the monitor.

d) Moreover, animals can lose consciousness to a certain degree if their brains work in different ways, either because of malfunction or sleep, in which case their ability to model the environment and/or themselves is reduced.

We have hence fulfilled requirements a)-d).

Present-day robots have low levels of consciousness, if any. They are able to perceive parts of their environment, and may be programmed to react to it, but their models are primitive and, so-far, mostly unpredictive. The same can be said about the present status of artificial intelligences. Deep learning, however, holds much promise because it allows a system to anticipate what is going to happen. Our definition implies that that to advance artificial intelligence, neural networks should preferably be fed with time-ordered sequences of events for only this will enable them to develop predictive models and, thereby, consciousness.

Based on the definition proposed above, we could, on principle attempt to scan all possible subsystems of a human brain, animals, or computers, and compute the fidelity of the subsystems with parts of the environment or other subsystems. This means, in principle we can answer Q1-Q3. However, even if it was possible to monitor a brain with sufficient accuracy this

task would presently be computationally infeasible. In the Appendix, I therefore propose how a presently possible practical test of the definition proposed here could look like.

What about the philosophical zombies? Can a system without experiences behave as if it was human? Yes, it can. The zombie's head could be empty except for a random generator that outputs arbitrary actions. By pure coincidence, the zombie might then behave exactly like a human and yet not have anything resembling experience according to our definition. This isn't impossible – but it is extremely unlikely. Indeed, as we have argued above, experience is hugely beneficial for the evolution of complex behavior that signals the system's ability to predict itself and its environment. Experience, hence, isn't necessary but likely to be found in any system able to behave even remotely human-like and philosophical zombies extremely rare.

In conclusion, we see that the trademark of consciousness is not that we think, but that we think about ourselves, about what others think, and what they think about what we may be thinking. We also, increasingly think about our own thinking, which might well be the path to higher consciousness.

Goals, Purpose, and Free Will

Let us then move towards derived concepts.

As previously noted, the ability of a system to predict the time-evolution of itself and the environment is limited because the more accurate a model, the more computationally intensive it becomes. This also means that generically a system can't predict its own reactions with absolute accuracy. This is why we have the impression of free will: Regardless of whether or not the future is indeed pre-determined, our inability to be sure about what we ourselves will do implies our self-models will project different possible future evolutions.

The trade-off between computational speed and accuracy also means that a system's ability in developing models will evolve towards a sweet spot where added computational effort for more accurate predictions is disfavored because the system's reaction would come too late to still bring more benefits.

Particularly valuable are hence models that have a low computational effort and yet are highly predictive. This is what goals are. We say that a system has a certain state (of itself and/or the subsystem) as "goal" if the optimization of actions to obtain the goal is predictive of the dynamics of the system.

Note that it isn't necessary for the system to actually reach the goal for it to be predictive! The goal of winning this essay content, for example, predicts that I agonize over every single word, regardless of whether I win.

We do not normally speak of 'goals' when referring to non-conscious systems, so to better match the common use of the word we could restricted this notion to conscious systems. Otherwise the moon could be said to have the goal of falling onto Earth, just that its initial velocity prevents it from ever reaching this goal. I leave it up to you whether or not you might want to add this restriction to the definition.

The existence of humans demonstrates that, given enough time, systems can develop remarkable predictive abilities and complicated reactions which in return are difficult to predict. Survival and reproduction, therefore, which once were highly predictive, might cease to be predictive in the long run. And so, when we start with assuming a goal to predict what happens in the course of evolution, we have it exactly the wrong way round: It is instead the possibility to make predictions with it that defines what we mean by goal.

Having come so far, it is straight-forward to also define what we mean by 'purpose.' We say an object or action has a certain 'purpose' if it increases the likelihood of a system reaching a goal. The word is also sometimes used to mean what we might call more descriptively 'intended purpose.' This is to mean a system might have computationally arrived at the conclusion that an action serves the purpose of reaching a goal, but the conclusion might be an inaccurate.

Humans are guilty of applying terms out of context, thus we often speak in extension of the purpose of abstract concepts other than objects or actions, for example the purpose of life. This is another big, bad question, to begin with because life, too, is not presently a well-defined notion. But even assuming that we had a good definition for life, speaking of its purpose would require us to first identify a goal that life might contribute to reaching.

Hydrogenating carbon dioxide, then, is not such a bad guess[8] for a goal that we could attribute to the universe, though it doesn't seem hugely predictive for the finer details. Based on the above we could instead conjecture that a more predictive goal for the evolution of the universe is becoming an accurate and predictive model of itself. The purpose of life, then, would be to develop these models. In other words, the purpose of our existence might be understanding the universe, but I admit that at this point I have succumbed to insubstantial speculation.

But we can now also ask what is the purpose of a specific life, yours or mine or the rabbit's. Let's take my life as example because it's the one I can speak about most confidently. To find the purpose of my life I must first ask which goal I am contributing to. Again, hydrogenating carbon dioxide rings true but is not very descriptive of me in particular. More predictive of my actions is that I try to increase the collective human understanding of the universe. The purpose of my life, hence, is to make others think. I hope I reached my goal.

References and Endnotes

[1] The origin of philosophical zombies seems to have gotten lost in the literature, but most relevant articles about it have been collected by David Chalmers and can be found at http://consc.net/online/1.3b, Retrieved March 3rd 2017

[2] Oizumi, M., Albantakis, L., Tononi, G., "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," PLoS Comput Biol. 10 (5): e1003588.

[3] Tegmark, M., "Consciousness as a State of Matter," Chaos Solitons Fractals 76, 238 (2015) [arXiv:1401.1219 [quant-ph]].

[4] Harvey, C. D., Collman, F., Dombeck, D. A., David W. Tank, D. W., "Intracellular dynamics of hippocampal place cells during virtual navigation," Nature 461, 941-946 (15 October 2009).

[5] Diester, I., Nieder, A. "Semantic Associations between Signs and Numerical Categories in the Prefrontal Cortex," PLoS Biol 5(11): e294 (2007).

[6] With apologies to the more physically inclined readers, I will in the following neglect special relativity and assume there is an absolute time by use of which we can speak of simultaneity so that it's unambiguous what it means for the morphism to be predictive. The definitions I propose here can be made compatible with special relativity by taking into account the finite time it takes information to travel through I/O channels but I space limitations prevent me from elaborating on this. Let me just say it can be done, but isn't relevant for what follows.

[7] Liu X, Ward BD, Binder JR, Li S-J, Hudetz AG (2014) Scale-Free Functional Connectivity of the Brain Is Maintained in Anesthetized Healthy Participants but Not in Patients with Unresponsive Wakefulness Syndrome. PLoS ONE 9(3): e92182.

[8] Attributed to Michael Russell in: Sean Carroll, "The Big Picture," Dutton (2016), p. 260

Appendix

A simple way to probe the consciousness-level of an anesthetized patient or a patient with locked-in syndrome might look as follows.

Level one: Testing for level one awareness is straight-forward. Play the patient a sound that repeats in regular intervals, or apply any other regular sensory stimulus, such as touch or, if necessary, direct brain stimulation. If the brain shows any reaction in the same regular sequence as the stimulus that constitutes a simple morphism. (Yes, it's hard to not at least be unconsciously aware of direct brain stimulation.)

Level two: If the patient is able to generate a predictive model, the brain signal should show additional activity if the sequential stimulus – after some period of repetition – changes or suddenly stops. We would be looking, essentially, for a sign of surprise which would demonstrate forecasting ability.

Testing awareness level three and four is much more difficult for it requires finding evidence for the integration of subsystems.

Level three: One way to demonstrate experience, according to our definition, would be to demonstrate that the patient is able to model not only the input signal but the connection between various types of input signals. This would show that the patient is – to some extent – aware of the way their own brain is connected internally and with the environment. This could

be done for example by probing whether the patient's reaction to two different kinds of stimulus is the same whether or not the input is synchronized, or where it is coming from.

Level four: Probing cognition isn't all that difficult because we are used to doing it. It can be done by looking for the ability of the patient to learn and react, for example by adding a feedback look to the input signal. A practical way to achieve this may be to continue a signal as long as the patient is displaying a certain brain activity. Over time, the patient should learn of this connection, meaning when the feedback is suddenly discontinued, there should again be a surprise reaction. The presently used tests, such as asking patients to imagine performing a certain action and measuring their response, are much more complex ways to test for level four awareness.