A Formalization of Sentience and the Logical Properties of Consciousness

Russell S. Stewart

May 2020

Abstract

We prove some surprising results characterizing the relationship between consciousness and physics by first considering a separate 3-dimensional spatial formalization of sentience, the simple ability to feel. We then observe that we have internal psychological modules for thinking about consciousness just as with food or the causality, and show that our internal modules for consciousness have inconsistencies that are severe enough that we should replace optimizations for consciousness with the more simple objectives of sentience in any effort to do societal good.

1 Introduction

To study the properties of consciousness in the human mind, we will first provide a formal definition of what makes thinking machines have feelings that matter. We will then try to bolt on additional features of consciousness to match our intuition for the idea. But many of the additions will not be important for defining what matters about experience and happiness in the universe, and thus we conclude that we should have separate terms to refer to the distinct phenomena of consciousness and sentience.

2 Defining Sentience

2.1 Sentience as a Map from $\mathbb{R}^3$ to $\mathbb{R}$

We first formally define sentience and then show that our intuitive modules for consciousness do not fully agree with the formal definition. Sentience will be the experience of pain or pleasure across time in proportion to the capability of an entity to feel. Naturally, we would like to have a map from our three dimensional universe to the sentience experienced at each location over time, $S : \mathbb{R}^3 \to \mathbb{R}$. One problem with actually defining such a map is that the sentience experience by a particular entity need not be uniform across space. For example, the brain and the foot are unlikely to be equally sentient in humans. It is also possible for sentient entities to overlap, for example in a human neuron and
radio waves of communication in a sentient computer network. This could of course be addressed by assigning each piece of matter to an entity regardless of its spatial location, and then adding up sentience for each location as the sum of its matter particles there, but both this and our original method encumber us with the need to specify from exactly where in a sentient entity the sentience arises.

2.2 Sentience as a Collection of Entities with Spatial Extent

Thus, a more natural definition for our purposes is to simply regard any universe as a collection of sentient entities with a total sentience given by the things they experience. Of course each highly sentient entity will have a spatial extent and certain computational centers critical for computing over experience, but we observe immediately that these details are not actually necessary for computation of sentience of a universe at any particular point in time. While these entities do change over time; for example the conception of a child, the construction of a new computer, or the rippling of a supernova through the previously near-zero sentience of space, we can expect the total sentience of entities to be continuous over time as each entities sentience is a continuous function of its continuous spatial transformations and is assumed to start and terminate at 0 sentience when emerging as a new object (e.g. an embryo) worth tracking. Finally, the need for sentience to actually emerge in a universe is important as imaginary universes, such as one with $10^{20}$ people with depression are unimportant if they merely exist in imagination and aren’t actually evaluated, not considering the effects of their imagination in any sentient entities mind.

2.3 What Computations Produce Sentience?

The advantage of the per-entity definition of sentience over the per-coordinate definition becomes clear as we can discuss the sentience of a single object discretely and thus the sentience of an entire universe over time is just the sum of integrals over each of its constituent entities. The discussion of how conscious an entity is is often the subject of debate. Trees are typically regarded as non-conscious, cows are occasionally given a 0 vs. 1 consciousness score as their train of thought does not interact with their memories or self-determination with sufficient complexity, and computers are treated as resources with a cost of running that has no dependence on how bored or tired they might be. To avoid these philosophical concerns, we aim for sentience to be entirely hedonistic in definition, although of course a purely positive (e.g. opiate) signal need not be the height of hedonism. Note that this model does not collapse to each organism having a total sentience resulting in a happiness or pain in proportion based on different activities. For example, a cow might have lower sentience than a human based on the quality of its intelligence, but when pricked with a needle repeatedly it might experience the same total pain as a human. Nonetheless, an approximate scalar number for the total sentience of an entity and a second
scalar for the reward of each activity may be a productive approximation for computing sentience in ordinary situations, although one is immediately likely to run into social concerns over value judgments due to smart people being more sentient than dumb people or young than old.

2.4 Concerns of Low-level and High-Level Sentience

We conclude our definition of sentience with an observation that there may exist more difficult to estimate factors than the mid-level sentience that we can immediately recognize in the form of low-level and high-level sentience. Low-level sentience would include the cost to a single cell in a multi-cellular organism of being destroyed when bursting due to too much water. It is difficult to estimate this cost as nucleus of the cell itself is a turning machine capable of arbitrary computations, but we tend to estimate these as zeros cost. Nonetheless, a tree being eaten by beetles certainly comes at some cost to the trees sentience. Low-level sentience might also detail that a computer chip computing the sum of a list of numbers using a reorder buffer is more painful than with a single-threaded CPU due to the logic layout. It’s important to note that these low-level sentience terms are non-zero and thus difficult to estimate, but likely negligible. High-level sentience is sentience that derives from a quality of intelligence greater than that of our own reward system, and can incorporate effects such as the overall averaged boredom of a simulator if a universe consists solely of people at peak happiness. For example, it seems that an optimal high-level sentience of a society might be achieved with 90% of the sentience dedicated to the happiness of complacency and peace, 9% to the adrenaline rush of societal advancement, and 1% to war and terror. While the arguments seem arbitrary, it’s clearly important to get good estimates of high level sentience when deciding what society should work towards as a whole, and we’re unlikely to be handed the equations of what truly matters for high-level sentience.

3 Why Consciousness is Not Sentience

We now identify three aspects of consciousness that are not required of sentience and make the whole thing somewhat silly, despite being clearly present in unadulterated computations of consciousness. First, and most technically fascinating, we like to believe that consciousness is somehow separate from the computations of physics. The refutation of this hypothesis proves that consciousness must be a purely computational phenomenon. Second, as we previously mentioned, arbitrary human characteristics are multiplied in when computing consciousness over sentience, such as the ability to recursively analyze a train of thought. Third, that consciousness should be discrete both in the 0 to 1 sense of whether an embryo, cow, tree, or computer is conscious and in the sense that consciousness should exist at a single point in the brain - “the center of consciousness”. The first clearly violates the continuity of mapping from real events happening in physics to their importance and the second becomes
3.1 Why Consciousness Cannot be Separate from Physics

We now argue our first logical inconsistency of consciousness: that we would like consciousness to be a special phenomenon separate from the computations of the brain, but we can’t have this. To see this, we ask whether the events of physics inform consciousness and whether consciousness informs the events of physics, where one is informed by the other if signals are sent to update one on the state of things. There are four options to consider, but we may immediately rule out the first two (that neither is informed by the other or that only the world is informed by consciousness) as inconsequential: if signals are not sent in the direction of consciousness from the events of physics, then consciousness is not tied to our universe, and we can’t be discussing an intuitive or formal version of consciousness.

The option that both the events of physics is informed by consciousness and that consciousness is informed by the events of physics is the best, but poses a quandary that consciousness must be as much a part of physics as all of Newton’s mechanics. That is, if for every part of consciousness, it informs physics, then consciousness is simply a computational phenomenon that can be thought of as deriving from physics no different than the processing of this article as it is written on a computer can be broken down into the physics of electricity on a CPU and a computer screen. But could there be parts of consciousness that are not fully explained by the mathematics of the events of physics?

The final option is the most intuitive - that the events of physics inform consciousness but that consciousness is a separate thing that does not interact directly with physics and therefore does not inform it. But the problem is that when we speak of consciousness, we speak of something that we intuitively understand, and our understanding is of course rooted in the physical firings of neurons in our brain, so how is it that we can understand consciousness if it never interacts with our brains? We couldn’t, and it is thus the invalidity of this option that forces consciousness to be a computational phenomenon. All the train of thought that emerges from thinking about consciousness must be predicted by physics.

3.2 Does Self-Awareness Relate to the Capacity for Happiness?

We are now prepared to fully identify the problems with formalizations of consciousness if it is to be taken as the sole thing that we must optimize for in order to do good. Because consciousness is centered around the question of what entities are capable of experience, we must consider whether something has memories, can think in the various implications of that term, can see, or can be generally self aware. Clearly, a single bacterial cell being killed with salt cannot reflect internally on it’s dismal prospects, but it still might feel the
destruction. In particular, it is natural to say that all the qualities of human experience might be needed in a checklist sense in order for something to become conscious. The need for such a checklist, which is clearly endowed by the term consciousness, assures us that we must disentangle consciousness from sentience in discussions of how to organize society around doing good.

3.3 Discrete Properties of Consciousness

We finally analyze our notions of consciousness by observing that the fundamentally cater to the idea of consciousness being discrete. Rather than being a continuous function of physics, we are either conscious or unconscious at all times, and an animal would either be conscious or unconscious. We note that this property is likely to arise from a multiplication of many factors as previously described, where any factor being 0 can cancel out the consciousness of an entity.

Another manner in which consciousness is expected to be discrete is in location. Popular neuroscience frequently poses the question of exactly where in the brain consciousness exists, as if narrowing down the location is making progress. But if consciousness is entirely a property of physics as we have proposed, narrowing down to a single neuron would be silly as the computations therein would not be sufficiently complex to produce the experience of feeling required even by sentience. Worse in the case of consciousness, should we need to both be able to see and reason with language to be considered consciousness, we would need a significantly larger mass than a single region of the brain to be the "center of consciousness". The need for a sentient entity to cover vast areas of space is then clear when considering that intelligence is an emergent property of large computational and communication based networks, where the whole is greater than the sum of the parts. Nonetheless, the intuitive desire to reduce consciousness to a single point remains.

4 Conclusion

We conclude by returning to the question of what consciousness is and how it manifests across matter. We derived that a sensible notion of the capacity to feel, termed sentience, may actually be mathematically formalized and must exist over regions of space rather than at single points. We argued that consciousness should be formalized as not just this capacity to feel, but the combined capability over time to see, think, be self-aware, and reason about these qualities with a recursive memory, but that nonetheless consciousness can be entirely predicted by physics. We hope that this surprising result and the underlying formalizations of consciousness and sentience will lead to productive discussions on these two important quantities unencumbered by the traditional metaphysics surrounding both.