The Substrate-Prior of Consciousness

Gabriel Leuenberger

2021[1]

1 Introduction

This essay is based on our previous draft paper [9] that addressed several mind-related philosophical questions by describing how their solutions could in principle be calculated. Since our paper showed itself to be difficult to understand, we now verbalized more of our reasoning and restricted this essay to a more basic question: Given functionally equivalent minds, how does the expected quantity of their conscious experience differ across different substrates[2] and how could we calculate this?

We argue that a realistic digital brain emulation would be orders of magnitude less conscious than a real biological brain. On the other hand, a mind running on neuromorphic hardware or a quantum computer could in principle be more conscious than than a biological brain.

2 Simulated Room Thought Experiment

Consider the following scenario: You are a broke philosopher and you meet a strange transhumanist brain scientist. He offers you a great monetary reward for participating in his experiment that uses a novel brain scanner. After agreeing to participate, you sign a statement that allows him to own all the information that is obtained from the scan. He lets you enter a special room while he waits in a different room. You notice that the room you are in only has a red door and a blue door.

You then receive a video call from the transhumanist. He explains to you that the scan was successfully completed and that he just used all of the detailed information about your brain cells in order to create a quite accurate simulation of your brain that is currently being run in real-time on a powerful supercomputer. This simulator is a complicated software that simulates membrane potentials of accurate virtual 3-D models of your brain cells. Not just your brain, but in fact your entire body and the room are being simulated with it.

Since you do not believe him, he then reconnects the video call such that you can talk to the simulated version of yourself. You then converse with this simulated version of yourself for a long time, until you are certain that the simulation is indeed functionally equivalent to yourself and has the same memories as you. The simulated version believes itself to be real and believes that you are simulated. You disagree and get into a long argument over this issue. Finally both of you have to concede that currently neither one of you knows for certain, who is the original and who is the copy.

After this insight of yours, the transhumanist rejoins the video call and explains that the promised monetary reward can be obtained by leaving through the red door; but if you leave through the blue door you will stay broke. As you start moving towards the red door, he goes on to explain that the simulation is set up, such that the simulated version of you will be terminated painlessly, without even noticing, when leaving through the blue door. However, if it leaves through the red door, it will encounter a torturous death. Since you do not really know whether you are being simulated or not, you back off from that red door.

[1] Written in June of 2020.

[2] The substrate is the type of physical device that contains a mind. Examples of substrata are biological brains and electronic processors or even Chinese rooms [18].

The transhumanist refuses to reveal whether you are the original or not, but he provides you with sufficient further information, to verify his claims about the two doors. He furthermore reveals that the torture would not be worth the money. You now have only two choices: leave through the red door or leave through the blue door. The room is set up such that it is impossible for you to obtain any empirical evidence that would reveal whether you are real or virtual. You now have to rely on your philosophy-skills to make this decision.

3 Solution via Universal Prior

In this section we find out whether you are the original or the copy and take the corresponding decision.

3.1 Occam's Razor and Minimum Description Length

The simulated room thought experiment leaves us with two explanations, two theories, or two hypotheses, or two models, that both equally well explain all of the observations that you are making while being in the room. In such cases one can try to apply one of the most fundamental principles of science and rationality, which is most commonly known as Occam's razor [4], although it was formulated as far back as Aristotle [3]. It is usually translated as: "entities should not be multiplied without necessity." , which is to be interpreted to mean that given multiple conflicting theories that equally well explain all of the observations made so far, one should choose the explanation that is the simplest, i.e.: the explanation that contains the least assumptions, or more precisely, the smallest amount of presupposed information. Note that the simplest theory does Not need to be the easiest one to find nor the easiest one to understand by any means. The principle was also restated by Newton [14], Einstein, and so on.

In the electronic computer age, several formalizations of Occam's razor were developed, collectively referred to as minimum description length (MDL) [1, 13, 17, 19]. The general idea is to select the shortest formal description or the shortest computer program that generates only all of the observed data, i.e. theories correspond to programs. The length of the program is measured by its information content, e.g.: By how many bits are required to write the program. The computer on which the program can run, is called the reference machine, which should itself be a most simple universal computer. Note that the shortest program's run-time does not need to be short at all. In the following subsections we start to apply MDL to the thought experiment.

3.2 Evidence for Digital Physics

Before moving to the solution, let us have a quick look at fundamental physics. It has been suspected for at least 50 years that our universe is a discrete mathematical structure that in principle could be entirely and accurately computed [21] by simple rules, that are simpler than contemporary theories consisting of definitions of continuous functions in continuous spaces with numeric constants. Fundamental physical limits to computation would be expected from digital physics, thus an example of evidence for digital physics is the existence of multiple ultimate physical limits to computation [11] (computation speed, energy requirements, memory) derived from contemporary physics. Further hints are the low-entropy initial state that is the big bang as well as the finiteness of the observable universe. Furthermore, general relativity can be re-derived using assumptions of digital physics [6, 12]. Hence, most recently, the Wolfram Physics Project was launched, aimed at finding the ultimate fundamental theory of physics that would be a program, which generates a hypergraph that is the universe [20], where the big bang is the start of the computation. Note that, presumably, given unlimited computational resources, such a program should generate every single Plank-scale event that ever happened in the history of our universe.

3.3 Observation and Subjective Conscious Experience

Before applying Occam's razor, we have to specify what your observation consists of. If we were to specify that your observation should consist only of a short verbal description of what the room looks like from the inside, we would miss out on a lot of additional information that you are perceiving. An alternative would be to specify your observation as consisting out of the data stream from your sensory nerves to your brain. However, for conscious rational decision making, this data stream is only available to you in a form that was already highly preprocessed within your brain. Furthermore this data stream does not include observations of your own memories and thoughts. The optimal solution would be to specify the observation as consisting of your entire subjective conscious experience and nothing more, since we do not want to leave out any information that you may know, nor include unwarranted assumptions about the outside. However, we can only use the aspects of your conscious experience that are describable in principle, such as a network of similarity-relations between colors, No indescribable aspects, such as the redness of red.

Representations of such conscious experience in the form of mathematical structures have been proposed for example by Tononi [15] and Goertzel [5]. Without contemplating whether their representations are exactly correct, we will work under the assumption that conscious experience can in principle be represented as a mathematical structure. This assumption becomes more plausible, if physics, at its most fundamental level, is fully computable, since conscious experience seems to be generated from physical events.

Unfortunately, you are unable to introspect and write down this mathematical structure, because it consists of too much information. Nevertheless, you could still take the rational decision without needing to analyze this mathematical structure directly, but instead, by reasoning about what you would decide if the mathematical structure were available, and then take that same decision. Therefore, for the remainder of this section, we can reason as if this mathematical structure were available to us. Let this mathematical structure of your conscious experience be called Q.

3.4 Multiple Components of the Shortest Program

Following MDL, we have to find the shortest possible program that generates Q. This shortest program represents the theory which contains the assumptions that are the most epistemically rational to believe, when knowing nothing but Q. Therefore, by introducing commonly held modern-day assumptions which we believe to be rational, we can infer what the properties of this shortest program may be. Our assumptions are the following:

Assumption 1: All physical events in the history of our universe are fully computable from a set of rules that is much simpler than a mind.

Assumption 2: Your mind is in some way located within this history of our universe that is vastly larger and older than your mind.

Assumption 3: The functioning of the mind is enabled by a large number of underlying physical events.

Assumption 4: The structure of conscious experience Q is a consequence of how the mind functions.

Therefore, by connecting these assumptions, we can infer that the shortest program contains the following components:

Component 4: Computes Q from a formal description μ of computations[3] that your mind performed according to its functional aspect (likely at neural level).

Component 3: Computes μ from the underlying fully detailed description of the physical events that occurred in your mind (from Plank-scale).

[3] Note here that 'computing Q' and the 'computations that your mind performed' are two different computations.

Component 2: Localisation: Searches the detailed history of our universe to find the spatiotemporal location of your mind according to simple search criteria. This component is inspired by Hutter's observer localisation [7], although he did not apply this idea to the mind.

Component 1: Computes the entire detailed history of our universe including all Planck-scale events, starting from the big bang.

To repeat the main claim here: These are the components of the shortest possible program that can generate the structure of your conscious experience Q. Let this program be called p_{min}.

Sidenote: If Q would itself consist of less information than a program consisting of such components, than the shortest program that describes Q would obviously not consist of the described components. This would happen if Q were the conscious experience of a small animal with very few neurons. This implies, that given only the small animals experience Q, it would not be rational to have the above assumptions. But since we assume the above assumptions to be rational, it implies that a human's Q is sufficiently large.

3.5 Comparing the two Substrates

Recall that we were left with two possibilities: Either you are the original, that inhabits the substrate consisting of biological neurons, or you are the copy, that inhabits the substrate consisting of software running on electronic processors. In order to choose between these two possibilities, we can compare the length that p_{min} would have, if you were located on either substrate, and then select the shorter one in accordance with Occam / MDL. We can conduct this length comparison by estimating the difference separately for the previously introduced four components of p_{min} and then adding up these differences to estimate the total difference in length, although it would be very difficult to obtain any accurate numbers. The following is the Comparison:

Component 4: Since the minds on the two substrates are functionally almost identical, component 4 must be almost identical for both cases and hence the expected difference in length is negligible.

Component 1: Since both substrates are located in the same universe, component 1 will be the same (or almost the same) for both substrates, and hence the length difference will be zero.

Component 3: Recall that this component has to compute μ, where μ is a formal description of computations that your mind performed according to its functional aspect, likely at the neural level. Component 3 computes μ from the underlying fully detailed description of the physical events that occurred in your mind. In order to accomplish this, for both, the substrate of biological neurons, as well as the substrate of software of electronic circuits, firstly the motions of electric charges have to be recognized. The main difference between the two substrates is the following: Recall that the simulator, running on a supercomputer, is a complicated software that simulates membrane potentials of accurate virtual 3-D models of your brain cells. For the original biological neural substrate, component 3 could already be relatively close to obtaining μ, just by having recognized the motion of electric charges. Not so for the digital substrate: After having recognized the motion of electric charges within computer circuits, a larger amount of additional information is required in order to interpret these many binary currents as numbers that somehow, hidden behind layers of software, represent coordinates, shapes, and electric activities of neurons, from which μ can ultimately be obtained. This required additional information causes a great difference between the substrates with regards to the length of component 3.

Component 2: The search criterion of the localisation would largely consist of information concerned with recognizing whether some matter is a mind of a certain substrate. To accomplish this recognition task, component 2 could make use of parts of component 3. But from the many recognized minds, still the right one (yours) needs to be selected. The additional amount of information required for this selection, only has to grow logarithmically w.r.t. the total number of minds, analogous to how the size of identification numbers only has to grow logarithmically w.r.t. the number of objects to be identified.

Therefore, between the substrates, we can assume the length difference in component 2 to be overshadowed by the length difference in component 3.

If we now add up the four differences, we should get a total length difference that is mostly impacted by component 3, which makes the length of p_{min} smaller in the case of the original substrate of biological neurons. Let this total length difference be called δ. Now you could use Occam's razor to shave away the possibility that you are simulated.

3.6 Epicurus' Multiple Explanations and Algorithmic Solomonoff Probability

As you are about to use Occam's razor to shave away the possibility that you are simulated, you suddenly remember Epicurus' principle of multiple explanations: "If several theories are consistent with the observed data, retain them all!". Indeed, you only found out that one of the two possibilities is more likely, but you have no idea how strong the odds are. The odds could be 51 to 49, in which case you would like to choose the blue door, since the torture is not worth the money. On the other hand, the odds could be 1000000 to 1, in which case you would like to choose the red door instead. So what is needed, before taking the decision, is a probability estimate, which we shall introduce here.

Epicurus' principle furthermore reminds us that there could not just be one theory supporting the claim that you are simulated, but there could instead be many such theories and therefore, being simulated could in principle be more likely, even though each one of these theories on its own is less likely. Solomonoff took such concerns seriously and developed the Universal Prior, also known as Algorithmic Probability [16, 19]. It is the probability distribution over all possible observations that are outputs of a simple universal computer, given that the program was selected uniformly at random from all possible programs. This means that the Universal Prior probability is obtained by considering all, infinitely many, possible theories simultaneously. However, remarkably, the Coding Theorem by Levin [10], showed that this Universal Prior probability can always be estimated by 2^{-K}, where K is the Algorithmic Kolmogorov Complexity [8] , that is the length (in bits) of the shortest program that outputs the observed data. Since, in our case, a rough estimate could be sufficient, we can use this formula, which only relies on the shortest, instead of all possible programs, thanks to the Coding Theorem.

3.7 The Odds and the Final Decision

The odds is a ratio between two probabilities. The odds we are interested in, is the ratio between the probabilities of your two possible substrates. As shown in the previous section, such probabilities can be estimated by 2^{-K}, where K is the length of p_{min} in bits, that differs between the two substrates. Therefore, by dividing one such probability by the other, we get the odds equal to 2^{δ}, where δ is the difference between the two K that we obtained in subsection 3.5.

If δ would be equal to a measly 20 bits, it would imply that you are about 2^{20} times more likely to be on one substrate over the other, i.e.: it would be about one million times more likely. Since we previously concluded that the value of δ is mostly influenced by the engineered complexity of the simulator, and you know software to usually be orders of magnitude more complex than 20 bits, you conclude that you must be many orders of magnitude more likely to be the real, original, philosopher consisting of biological neurons. You finally leave through the red door and happily take the money.

Luckily it was not some efficient brain emulation running on specialized neuromorphic hardware without complex software, otherwise this decision would have been more tricky.

4 Discussion

4.1 Substrate Prior of Consciousness

If you would not know where you are, but you only knew that there are a thousand people at place A and only ten people at place B, then you would assume that you are about a hundred times more likely to be at place A. The total amount of consciousness at place A is also hundred times greater than at place B. It becomes obvious that such probabilities scale with the quantities of consciousness, or, more generally, with the expected quantities of consciousness.

In the previous section we have seen that, if you are a conscious mind, a priori, the probability of being a certain mind can differ drastically solely based on the properties of the substrate, despite functional equivalence. This means that we can define a prior probability distribution over substrates; let it be called the Substrate Prior of Consciousness. Given a set of substrates, an estimate of this prior could in principle be calculated by following the steps of the previous section. However, presently achieving accuracy is not possible because the ultimate theory of fundamental physics remained unknown to this day, as well as how to represent Q is unknown, as well as how exactly the human brain functions is unknown, all of which would be required for an accurate calculation, and all of which could potentially become known in the following decades.

This Substrate Prior of Consciousness could be interpreted as an expectation value of the quantity of consciousness of a mind on a substrate relative to other substrates. A number that is an expectation value of the quantity of consciousness can then be interpreted in the following two major ways:

1. This number is the probability of being conscious as opposed to being unconscious.

2. This number is the quantity of consciousness itself.

We advocate the second interpretation, since it avoids the strange case of a person that believes itself to be conscious but is in fact completely unconscious; a type of philosophical zombie. However there can easily be extremely low substrate priors and therefore there can still be persons that have such a low quantity of consciousness that we could consider them philosophical zombies in practice.

Note that this quantity of consciousness should not be confounded with the degree of wakefulness, since the word conscious can mean 'being awake', as opposed to 'being in deep sleep', unconscious.

4.2 Importance for Mind Uploading and the Ethics of AI

Unfortunately, the most technologically feasible way to upload your mind into a computer would be to have a machine repeatedly slice small pieces off your brain in order to scan them with sufficient detail, until your brain is completely disassembled. This is called destructive mind uploading, since the original mind is destroyed. If you were an old person with a disease that is slowly deteriorating your mental health and your life quality, You might want to use this technology in order to continue your life as a cyborg without the disease. Such a decision could depend on the substrate prior, since, if the computer's substrate prior would turn out to be very low, your electronic afterlife would be worthless compared to your current life, which you hence would decide to retain.

Conversely, given the technology of non-destructive mind uploading, in order to complete an overwhelming amount of annoying work, you could make numerous virtual copies of yourself and enslave them to complete this work. The decision to do this, again, could hinge on the substrate prior, since, if the computer's substrate prior would turn out to be relatively high, the virtual self-enslavement could potentially inflict great suffering on yourselves, which you therefore would want to avoid. If, on the other hand, the computer's substrate prior would turn out to be relatively low, you would not have to worry about fully exploiting these virtual minds.

The same reasoning can be applied to the ethics of artificial intelligence in the question regarding the exploitation of generally intelligent machines. Bostrom and Yudkowsky introduced the 'Principle of Substrate Non-Discrimination' [2], which reads: "If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status." Here "same conscious experience" should probably be replaced by "same quantity of conscious experience"; a condition, which, as we have seen, is not to be taken as a given.

The method presented in our essay can be applied to a wide range of other related philosophical problems, to be revealed in future publications.

References

[1] Lloyd Allison. *Coding Ockham's Razor*. Springer, 2018.

[2] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1:316–334, 2014.

[3] MJ Charlesworth. Aristotle's razor. *Philosophical Studies*, 6:105–112, 1955.

[4] Guilelmus de Ockham et al. *Philosophical writings: a selection*. Hackett Publishing, 1990.

[5] Ben Goertzel. Hyperset models of self, will and reflective consciousness. *International Journal of Machine Consciousness*, 3(01):19–53, 2011.

[6] Jonathan Gorard. Some relativistic and gravitational properties of the wolfram model. *arXiv preprint arXiv:2004.14810*, 2020.

[7] Marcus Hutter. The subjective computable universe. In *A Computable Universe: Understanding and Exploring Nature as Computation*, pages 399–416. World Scientific, 2013.

[8] Andrei Nikolaevich Kolmogorov. Three approaches to the quantitative definition of information. *International journal of computer mathematics*, 2(1-4):157–168, 1968.

[9] Gabriel Leuenberger. Applications of algorithmic probability to the philosophy of mind. *arXiv preprint arXiv:1404.1718*, 2014.

[10] Leonid Anatolevich Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informatsii*, 10(3):30–35, 1974.

[11] Seth Lloyd. Ultimate physical limits to computation. *Nature*, 406(6799):1047–1054, 2000.

[12] Seth Lloyd. The computational universe: quantum gravity from quantum computation. Technical report, 2005.

[13] Volker Nannen. A short introduction to model selection, kolmogorov complexity and minimum description length (mdl). *arXiv preprint arXiv:1005.2364*, 2010.

[14] Isaac Newton. *The Principia: mathematical principles of natural philosophy*. Univ of California Press, 1999. See Rule I of Book III.

[15] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol*, 10(5):e1003588, 2014.

[16] Samuel Rathmanner and Marcus Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136, 2011.

[17] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[18] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[19] Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

[20] Stephen Wolfram. A class of models with the potential to represent fundamental physics. *arXiv*, pages arXiv–2004, 2020.

[21] Konrad Zuse. Rechnender raum: Schriften zur datenverarbeitung. *Vieweg, Braunschweig*, 42, 1969.