What Is It Like to Be a Brain Simulation?

Eray Özkural

Gök Us Sibernetik Araştırma ve Geliştirme Ltd. Şti.

Abstract. We frame the question of what kind of subjective experience a brain simulation would have in contrast to a biological brain. We discuss the brain prosthesis thought experiment. We evaluate how the experience of the brain simulation might differ from the biological, according to a number of hypotheses about experience and the properties of simulation. Then, we identify finer questions relating to the original inquiry, and answer them from both a general physicalist, and panexperientialist perspective.

1 Introduction

The nature of experience is one of those deep philosophical questions which philosophers and scientists alike have not been able to reach a consensus on. In this article, I review a computational variant of a basic question of *subjectivity*. In his classical article "What is it like to be a bat?", Thomas Nagel investigates whether we can give a satisfactory answer to the question in the title of his article, and due to what he thinks to be fundamental barriers, concludes that it is not something we humans can know [1]. We can intuitively agree that although the bat's brain must have many similarities to a human's, since both species are mammalian, the bat brain contains a sensory modality quite unlike any which we possess. By induction, we can guess that perhaps the difference between sonar perception and our visual experience could be as much as the difference between our visual and auditory perception. Yet, in some sense sonar is both visual and auditory, and still it is neither visual nor auditory. It is similar to vision, because it helps build a model of the scene around us, however, instead of stereoscopic vision, the bat sonar can make accurate 3-D models of the environment from a particular point of view, in contrast with normal vision that is said to have "2-1/2D vision" – it may also be contasted with blind people using audio and tactile perceptions. It is unlike anything that humans experience, and perhaps our wildest imaginations of bat sonar experience are doomed to fall short of the real thing. Namely, because it is difficult for us to understand the experience of a detailed and rapidly updated 3-D scene that does not contain optical experience as there is no 2-D image data from eyes to be interpreted. This would likely require specialized neural circuitry. And despite what Nagel has in mind, it seems theoretically possible to "download" bat sonar circuitry into a human brain (by growing the required neural module according to a given specification, connected to sonar equipment implanted in the body) so that the human can experience

the same sensory modality. In this problem, armchair philosophy alone may not be sufficient. The barrier to knowing what it is like to be a bat is, thus, mostly a technological barrier, not a conceptual or fundamental barrier, although, ultimately we cannot expect one to know *exactly* what a bat experiences, short of being one. In the best case, we would know what a bat experience is like, as the human brain could be augmented with a reconstruction of the perceptual brain circuit.
That being the case, we may also consider what a brain simulation, or an "upload" as affectionately called in science fiction literature, would experience, or whether it would experience anything at all, as brain simulation is a primary research goal on which computational neuroscientists have already made progress, e.g., [2]. The question that I pose is harder because the so-called upload does not run on a biological nervous system, and it is easier because the computation is the simulation of a human brain and not the biological computation of a bat brain, which is harder because of sonar perception. Answering this question is important, because presumably the subjective experience, raw sensations and feelings of a functional human brain are very personal and valuable to human beings. We would like to know if there is a substantial loss or difference in the quality of experience for our digital progeny. A recent survey of large-scale brain simulation projects may be found in [3].

2 Brain prosthesis thought experiment

The question is quite similar to the brain prosthesis thought experiment, in which biological neurons of a brain are gradually replaced by functionally equivalent (same input/output behavior) synthetic (electronic) neurons [4]. In that thought experiment, we ponder how the subjective experience of the brain would change. Although there are challenging problems such as interfacing smoothly with existing neural tissue, it is a scientifically plausible thought experiment, also discussed at some length in [5, Section 26.4]. Moravec suggests that nothing would change with respect to conscious experience in his book. Marvin Minsky has written similarly while discussing whether machines can be conscious [6]. He produces an argument similar to Wittgenstein's beetle-in-a-box thought experiment: since a brain simulation is supposed to be functionally equivalent, its utterances would be complete, and the brain simulation would know consciousness and claim to be conscious; why should we think that the simulation is lying deliberately? This is a convincing argument, however, it neglects to mention that subjective experience may not be identical to conscious cognition as usually assumed.
Contrariwise, John R. Searle maintains that the experience would gradually vanish in his book titled "The Rediscovery of the Mind" [7]. The reasoning of Minsky and Moravec seems to be that it is sufficient for the entire neural computation to be equivalent at the level of electrical signaling (as the synthetic neurons are electronic), while they seem to disregard other brain states. While for Searle, experience can only exist in "the right stuff", which he seems to be taking as biological substrate, although one cannot be certain [8]. We will revisit this division of views, for we shall identify yet another possibility.

3 The debate

Let us now frame the debate more thoroughly, given our small excursion to the origin of the thought experiment. On one side, AI researchers like Minsky and Moravec seem to think that simulating a brain will just work, and experience will be unchanged. On the other side, skeptics like Searle and Penrose, try everything to deny "consciousness" to poor machinekind. Although both Searle and Penrose are purportedly physicalists, they do not refrain from seeking almost magical events to explain experience.
However, it is not likely that word play will aid us much. We need to have a good scientific theory of when and how experience occurs. The best theory will have to be induced from experimental neuroscience and related facts. What is the most basic criterion for assessing whether the theory of experience is scientifically sound? No doubt, it comes down to rejecting every kind of superstitious explanation and approach this matter the same way as we are investigating problems in molecular biology, that subjective experience is ultimately made up of physical resources and interactions, and there is nothing else to it; this is a view also held by Minsky as he likens mysticism regarding consciousness to vitalism [6]. In philosophy, this approach to mind is called physicalism. A popular statement of physicalism is *token physicalism*: every mental event x is identical to a physical event y. That is a general hypothesis that neuroscientists already accept, because presumably, when the neuroscientist introduces a change to the brain, he would expect a corresponding change in the mental state, and he would expect that he can decode mental states from fMRI scans of the visual cortex as in several experiments. One may think of cybernetic eye implants and transcranial magnetic stimulation and confirm that this holds in practice, and that the hypothesis is scientifically plausible, for counter-examples are practically impossible to find. Another popular formulation of physicalism is the psychophysical identity theory [9]: that every experience is identical with some physical state. We accept both formulations at once, because the physicalist position is empirically supported, while metaphysical positions like predicate dualism are not.

4 Asking the question in the right way

We have discussed every basic concept to frame the question in a way akin to analysis. Mental events/states are physical events/states. Some neural events of a man *constitute* his subjective experience. The question is whether a whole brain simulation will have experience, and if it does, how similar this experience is to the experience of a human being. If the proponents of pan-experientialism are right, then this is nothing special, it is a basic capability of every physical resource (per the scientifically plausible, physicalist variant of pan-experientialism). However, we may question what physical states are part of human experience. We do not usually think that, for instance, a mitochondrial function inside neurons, or DNA, is part of the experience of the nervous system, because they

do not seem to be directly participating in the main function of the nervous system: thinking. They are not part of the causal picture of thought. Likewise, we do not assume that the power supply is part of the computation in a computer.
This analogy might seem out of place, initially. If pan-experientialists are right, experience is one of the basic features of the universe. It would then be all around us, however, most of it would *not* be organized as an intelligent mechanism, and therefore, correctly, we do not call them conscious. The claim that any physical system yields experience anywhere, is the simplest possible explanation of experience that is consistent with experiment, therefore it is a likely scientific hypothesis. It does not require any special or strange posits, conscious experience would then require merely physical resources organized in the right way so as to yield an intelligent functional mind. Consider my "evil alien" thought experiment. If tonight, an evil alien arrived and during your sleep shuffled all the connections in your brain randomly, would you still be intelligent? Very unlikely, since the connection pattern determines your brain function. You would lose all of your cognition, intelligence and memory. However, one is forced to accept that even in that state, one would likely have an experience, an experience that is probably *meaningless* and *chaotic*, but an experience nonetheless. Perhaps, that is what a glob of plasma experiences. The evil alien thought experiment supports the distinction between experience and consciousness. Many philosophers mistakenly think that consciousness consists in experience. That, when we understand the "mystery" of experience, we will understand consciousness. However, this is not the case. Experience is part of human-like consciousness, indeed, however, consciousness also includes a number of high-level cognitive functions such as reasoning, prediction, perception, awareness, self-reflection and so forth [10, Section 4]. I suggest that it is a valid hypothesis that there are entities that have experience without any recognizable mentality.

5 Neural code vs. neural states

Consider the hypothesis that experience is determined by particular neural codes. If that is true, even the experience of two humans is very different, because it has been shown that neural codes evolve in different ways [11]. One cannot simply substitute the code from a human for the code in someone else's brain, it will be random to the second human. And if the hypothesis is true, it will be another kind of experience, which basically means that the blue that I experience is different from the blue that you experience, while some assume we have no way of directly comparing them. Strange as that may sound, as it is based on sound neuroscience research, it is a point of view we must take seriously.
Yet even if the experiences of two humans can be very different, they must be sharing some basic fabric or property of experience. Where does that come from? If experience is this complex time evolution of electro-chemical signals, then it is in the shared nature of these electro-chemical signals and their processing that provides the computational platform. Remember that a change in the neural code (spike train) implies a lot of

changes. First of all, the chemical transmission across chemical synapses would change.Therefore, even a brain prosthesis device that simulates all the electrical signaling extremely accurately, might still miss part of the experience, if the bio-chemical events that occur in the brain are part of experience. Second, the electro-magnetic (EM) fields would change. Third, the computation would change (since data changes), although the basic "firmware" (genetic code) of the nervous system usually does not change.

To answer the question decisively, we must first encourage the neuroscientists to attack the problem of human experience, and find the sufficient and necessary conditions for experience to occur, or be transplanted from one person to the other. They should also find to what extent chemical reactions or other physical events are part of experience. It seems that chemical states may turn out to be important, and if as some people hypothesize quantum phenomena play a role in the brain, it may even be possible that the quantum descriptions may be relevant. If, for instance, we discover that the distinctive properties of nervous system experience crucially depend on quantum computations carried out at synapses and inside neurons, to construct the same kind of experience you would need similar physics and method of computation rather than a conventional electronic computer (a hypothesis also suggested in [12]). There is evidence that biology may exploit quantum computation though, i.e., recent experiments suggest that quantum coherence plays a key role in photosynthesis [13].

On the other hand, we may consider the minimalist hypothesis that electronic motion patterns may be a crucial part of experience, due to the energy and information they encompass, so perhaps electronic devices already contain brain-like experience. Then, the precise geometry and connectivity of the electronic circuit would be significant. This is much different from Searle, since we know that electrical signaling is a specific physical mechanism that plays a role in neural processing, and we do not assume that electrons have uncomputable, incomprehensible causal powers as Searle grants to biological stuff. A more intuitive possibility is that electromagnetic (EM) fields generated in the brain are the basis of experience, in which case the topology, amplitude, timing and other properties of electrical signaling may be relevant, i.e., anything that would change the EM field. EM theories of experience have been previously proposed, e.g., [14].

6 Simulation and transcoding experience

At this point, the reader might be wondering if the subject were not simulation: is the question like whether the simulation of rain is wet? In some respects, it is, because obviously, the simulation of water on a digital computer is not wet in the ordinary sense. Even a universal quantum computer [15] would not produce any real wetness, and all properties of water such as wetness are wholly composed of quantum mechanical properties – it is neither magic, nor an illusion. We may reconsider the question of experience of a brain simulation. We have a human brain A,

a joyous lump of meat, and its digitized form B, running on a digital computer. Will B's experience be the same as A's, or different, or non-existent? Up to now, if we accept the simplest theory of experience (that it requires no special conditions to exist at all), then we conclude that B will have some experience, but since the physical material is different, it will have a different texture to it. Otherwise, an accurate simulation, by definition, stores the same functional organization of cognitive constructs, like perception, memory, prediction, reflexes, emotions without significant information loss, and since the oft-dreaded panpsychism may be considered possible, they might give rise to an experience somewhat similar to the human brain, yet the computer program B, may be experiencing something else at the very lowest level. Simply because it is running on some future nanoprocessor instead of the brain, the physical states have become altogether different, yet their relative relationship, i.e., the *logical structure* of experience, is preserved.

Let us try to present the idea more intuitively. The brain is some kind of an analog/biological computer. A memorable analogy is the transfer of a 35mm film to a digital format. Surely, many critics have held that the digital format will be ultimately inferior, and indeed the medium and method of information storage is altogether different but the digital medium has its affordances like being able to backup and copy easily. In both formats, the "same information" is stored, yet the medium varies – in reality, there is no abstract object as information, only physical codes, thus "same information" just means bi-directional translatability of codes. Likewise, B's experience will have a different physical texture but its organization can be similar, even if the code of the simulation program of B will necessarily introduce significant physical difference – for instance neural signals may be represented by a binary code rather than a temporal analog signal. Perhaps, the atoms and thus, the fabric of B's experience will be different altogether as they are made up of the physical instances of computer code running on a digital computer. As improbable as it may seem today, these simulated minds will be made up of live computer codes, so it would be naive to expect that their nature will be the same as ours. They are not human brains, they are bio-information based artificial intelligences. In all likelihood, our experience would necessarily involve a degree of unimaginable features for them, as they are forced to simulate our physical make-up in their own computational architecture. This brings a degree of relative dissimilarity. And other physical differences only amplify this difference. Assuming the above explanation, therefore, when they are viewing the same scene, both A and B will claim to be experiencing the scene as they always did, and they will additionally claim that no change has occurred since the non-destructive uploading operation went successfully. This will be the case, because the state of experience is best understood as a feature of short-term memory, which has a distributed volatile memory architecture. There is a complex electro-chemical state that is held in memory with some effort, by making the same synapses repeat firing consistently, so that more or less the same physical state is maintained. This is what must be happening when you remember something, a neural state that is somewhat similar to when the event happened should be invoked. Since

in B, the fabric has changed, the memory will be reenacted in a different fabric, and therefore B will have no memory of what it used to feel like being A. Within the general framework of physicalism, we can claim that further significant changes will also influence B's experience. For instance, it will change execution to work on hardware with less communication latency or network topology. Or perhaps if the simulation is running on a different kind of architecture (like a PC), then the physical relations may change (such as time and geometry) and this may influence B's experience further. We can imagine this to be asking what happens when we simulate a complex 3-D computer architecture on a 2-D chip. We must maintain, however, that strict physicalism leads us to reject the idea that no mental changes happen when significant physical changes happen. If that were possible, then we would have to reject the idea that mental states are identical to physical states, which would be dualism. Moreover, a precise answer seems to depend on a number of smaller questions that we have little knowledge or certainty of. Some questions in this vein may be framed as: *Question 1:* What is the right level of simulation for B to be functionally equivalent to A? *Question 2:* How can the ontological contribution of the medium to experience be quantified? *Question 3:* Does experience crucially depend on any uncanny physics like quantum coherence?

6.1 General physicalist perspective

At this point, since we do not have conclusive scientific evidence, this is merely guesswork, and I shall give conservative answers. *Question 1*: If certain bio-chemical interactions are essential for the functions of emotions and sensations (like pleasure), then not simulating them adequately would result in a definite loss of functional accuracy. B would not work the same way, behaviorally, as A. This is true even if spike trains and changes in neural organization (plasticity) are simulated accurately otherwise. It is also not known with certainty whether we can simulate at a higher level, for instance via Artificial Neural Networks, that have abstracted the physiological characteristics altogether and just use numbers and arrows to represent A, or use mathematical abstractions to represent larger circuits. A recent brain simulation work shows that this might be possible [6]. It is important to know these so that B does not lack some significant cognitive functions of A, such as emotions. The right level of simulation seems to be at the level of molecular interactions which would at least cover the differences among various neurotransmitters, and which we can simulate on digital computers (perhaps imprecisely, though). At least this would be necessary because we know that, for instance, neurotransmitter levels and distribution influence behavior. Thus, it would be prudent to be able to accurately simulate the neurologically relevant biochemistry and dynamics of the brain, without necessarily simulating genetics or cell operation. *Question 2*: The most general characterizations may use information theory [16] or quantum information theory to quantify the amount of experience a system provides, and dissimilarity with another. An appropriate physical and informational framework must be chosen to answer this question in a satisfactory manner.

We can claim that ultimately low-level physical states must be part of experience, because there is no good alternative. The only alternative would be dualism, which is unacceptable to a physicalist. For a general physicalist, accepting a strong form of physicalism (that every mental event/property/predicate is physical), it seems prudent to think that the medium contributes to experience insofar as it influences computational states relevant to cognition, most significantly short-term memory. Thus, physicalism may force us to consider the hypothesis that physical details of both electrical and chemical neural events would be significant. In other words, a good deal of neurophysics could be included, there may be no simple answer as panexperientialists hope. It is likely that the atoms of experience belong to a specific physical kind, such as an EM field, or quantum superposition states, which may simplify quantification. The correct theory would likely give a measure of complexity and distinguish blue experience from green experience on that basis, reducing the difference to fundamental physical distinctions. *Question 3*: Some opponents of AI, most notably Penrose, have held that consciousness is due to macroscopic quantum phenomena (like laser) together with Hameroff [17], by which they try to explain unity of experience. While on the other hand, many philosophers of AI think that the unity is an illusion [18]. Yet, the illusion is something to explain, and it may well be that certain quantum interactions may be necessary for experience to occur, much like superconductivity. This again seems to be a scientific hypothesis, which can be tested. For a physicalist, thus, this is an unsettled matter, open to future research.

6.2 Panexperientialist perspective

An often underrated theory of experience is panpsychism, the view that all matter has mental properties. It is falsely believed by some that panpsychism is necessarily incompatible with physicalism. However, this is far from a settled controversy. Strawson has recently claimed that physicalism *entails* panpsychism [19]. More plausible is the view of panexperientialism: that experience resides in every physical system, however, not everything is a conscious mind, for that requires *cognition* in addition. Panpsychism is also proposed as an admissible philosophical interpretation of human-like AI experience in [20].The evidence from psychedelic drugs and anesthesia imply that changing the brain chemistry modulates experience. If the experience changes, what can this be attributed to? Does the basic computation change, or are chemical/quantum interactions actually part of human experience? It seems that panexperientialism is indeed the simplest theory of experience that is consistent with our observations, i.e., that every physical system may have the potential for conscious experience. Assume that the theory is right. Then, when we ask a physicist to quantify that, she may want to measure the energy, or the amount of computation or communication, or information content, or heat, whichever works the best. A general characterization of experience such that it would hold for any physical system, may be defined precisely, and may be part of experiments. It

would seem to me that the best characterization then would use information theory, because experience would not matter if it did not contain any information. For instance, an experience without any information could not contain any pictures or words. I suggest that we use such methods to clarify these finer questions. Also, the slightly more complex EM field theory has better empirical support (e.g., complexity of EM field rises with conscious thought, transcranial magnetic stimulation works), so it may be considered more restricted and more probable than general panexperientialism. Assuming the physicalist version of panexperientialism I may attempt to refine the answers above. *Question 1*: The first question is not dependent on experience, it is rather a question of which processes must be simulated for correct operation, so the answer does not change. *Question 2*: The biological medium seems to contribute at least as much as required for correct functionality (i.e., corresponding to neural information processing and biochemical changes precisely), and at most all the information as present in the biological biochemistry (i.e., precise cellular simulations), if we subscribe to panexperientialism. Co-located physical events might be significant in addition to electrical signals. According to the most general kind of panexperientialism, the cellular experience might simply constitute the low level texture of the collective experience of neural cell assemblies. Information integration theory of qualia [16] is a sort of panexperientialism. An EM field theory of experience suggests that only EM fields have experience which is a restricted kind of panexperientialism. Likewise with quantum computation hypothesis, which would imply that every varying make of quantum computer may yield different subjective experience. *Question 3*: Not necessarily. According to panexperientialism, it may be claimed to be likely false, since it would constrain minds to uncanny physics. If, for instance, quantum coherence is indeed prevalent in the brain and provides the experiential states, then the panexperientialist could point out to the possibility of a universal wave function (following the Many Worlds Interpretation). Another possibility is the use of a general physical theory, such as relativity or string theory to describe the body of experience.

Acknowledgements

Thanks to Ben Goertzel for his detailed comments that improved the article considerably. Thanks to anonymous reviewers, Joseph Polanik, and Peter D. Jones for their comments on the article.

References

1. Nagel, T.: What is it like to be a bat? Philosophical Review (1974)
2. Izhikevich, E.M., Edelman, G.M.: Large-scale model of mammalian thalamocortical systems. Proceedings of the National Academy of Sciences of the United States of America **105**(9) (2008) 3593–3598

3. Garis, H.D., Shuo, C., Goertzel, B., Ruiting, L.: A world survey of artificial brain projects, part i: Large-scale brain simulations. Neurocomputing **74**(1-3) (2010) 3–29
4. Moravec, H.: Mind Children: The Future of Robot and Human Intelligence. Harvard University Press (1990)
5. Russell, S., Norvig, P.: Artificial Intelligence A Modern Approach. Prentice-Hall Int. (1995)
6. Minsky, M.: Conscious machines. In: "Machinery of Consciousness", Proceedings, National Research Council of Canada, 75th Anniversary Symposium on Science in Society. (June 1991)
7. Searle, J.R.: The Rediscovery of the Mind. Bradford (1992)
8. Searle, J.: Minds, brains, and programs. Behavioral and Brain Sciences (1980)
9. Lewis, D.: An argument for the identity theory. Journal of Philosophy **63**(2) (1966) 17–25
10. Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster (2006)
11. Schneidman, E., Brenner, N., Tishby, N., de Ruyter van Steveninck, R., Bialek, W.: Universality and individuality in a neural code. In: Advances in Neural Information Processing 13, MIT Press (2001) 159–165
12. Goertzel, B.: 11 Consciousness. In: The Structure of Intelligence: A New Mathematical Model of Mind. Springer-Verlag (1993)
13. Panitchayangkoon, G., Voronine, D.V., Abramavicius, D., Caram, J.R., Lewis, N.H.C., Mukamel, S., Engel, G.S.: Direct evidence of quantum transport in photosynthetic light-harvesting complexes. Proceedings of the National Academy of Sciences **108**(52) (2011) 20908–20912
14. McFadden, J.: The conscious electromagnetic information (cemi) field theory: The hard problem made easy? Journal of Consciousness Studies **9**(8) (2002) 45–60
15. Lloyd, S.: Universal quantum simulators. Science **273**(5278) (1996) 1073–1078
16. Tononi, G.: Consciousness, information integration, and the brain. Progress in Brain Research **150** (2005) 109–126
17. Hameroff, S., Penrose, R.: Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness. Math. Comput. Simul. **40** (April 1996) 453–480
18. Minsky, M.: Decentralized minds. Behavioral and Brain Sciences **3**(03) (1980) 439–440
19. Strawson, G.: Realistic monism - why physicalism entails panpsychism. Journal of Consciousness Studies **13**(10-11) (2006) 3–31
20. Goertzel, B.: Hyperset models of self, will and reflective consciousness. International Journal of Machine Consciousness **3**(1) (June 2011)