Borderless Philosophy 3 (2020): 199-211.

On Ethical Problems of Whole Brain Simulations

Andreas Keller

I. Introduction

In 2012, *Scientific American* published an article written by the neuroscientist Henry Markram about "The Human Brain Project," a so-called "flagship project" of the European Union.[1]

The project, as originally planned by Markram, received a lot of criticism and was eventually reorganized and its focus changed, with more realistic aims compared to the original project.[2] However, I am interested here in philosophical and especially ethical issues about the original project, as presented by Markram in his article, since it can serve as an interesting thought experiment even if it might have been overambitious at its time.[3] But before I start the discussion, let me first give a short overview of Markram's original article.

II. The Human Brain Project

In his article, Markram develops the idea of simulating a whole human brain on supercomputers, as a research tool for scientists. He writes:

> A digital brain will be a resource for the entire scientific community: researchers will reserve time on it, as they do on the biggest telescopes, to conduct their experiments. They will use it to test theories of how the human brain works in health and in disease. They will recruit it to help them develop not only new diagnostic tests for

---

[1] H. Markram, "A Countdown to a Digital Simulation of Every Last Neuron in the Human Brain," *Scientific American* 306 (2012): 34-39, also available online at URL = <https://www.scientificamerican.com/article/human-brain-project-digital-simulation-neuron/>.

[2] See, e.g., S. Theil, "Why the Human Brain Project Went Wrong—and How to Fix It," *Scientific American* 313 (2015): 36-42, also available online at URL = <<https://www.scientificamerican.com/article/why-the-human-brain-project-went-wrong-and-how-to-fix-it/>. You can find a lot of material on the web and in the media about the project and about its eventual reorganization, but I am not going to go into that material here.

[3] This article is an extended and revised version of a letter to the editors of Scientific American in June 2012 containing my thoughts about Markram's article. They did not print it, and it was obviously too long and too complex to be easily abridged to be considered for publication. In November of the same year, I published a slightly revised version of this letter on one of my blogs under the title "An Open Letter to The Human Brain Project." I then sent an email with a link to my blog article to The Human Brain Project and another email to the authorities of the European Union responsible for funding the project. I did not receive an answer from members of the project, although I received an answer from somebody in the EU bureaucracy. There are other, similar projects, and I am using this one just as an example.

> autism or schizophrenia but also new therapies for depression and Alzheimer's disease. The wiring plan for tens of trillions of neural circuits will inspire the design of brainlike computers and intelligent robots. In short, it will transform neuroscience, medicine and information technology.[4]

Since it is probably impossible to measure the exact way the neuron in the brain are connected, Markram's approach was to let those connections develop by themselves like they do in a fetus:

> The key to our approach is to craft the basic blueprint according to which the brain is built: the set of rules that has guided its construction over evolution and does so anew in each developing fetus. In theory, those rules are all the information we need to start building a brain.[5]

Markram envisions a brain simulation with several layers: "molecular," "cellular," "circuits," "regions," and "whole organ." According to Markram's conception, we would reach the computing power required for a whole brain simulation by about 2023.

In this way, Markram describes the project of simulating a whole human brain in a computer system, down to the molecular level. He argues that technical advances over the next decades might make this feasible. The project, as described in the article, might be criticized in several ways. It was overambitious for its time. For example, the simulation of neurons on the molecular level might turn out to be far more difficult than Markram presents it here, for reasons of computational complexity alone.[6]

I am not going to discuss here how realistic Markram's project was or whether what he presents in that article is a sound and reputable scientific approach.[7] Instead, on philosophico-ethical grounds, I am mainly interested in the following paragraphs:

> The first whole-brain simulations we run on our instrument will lack a fundamental feature of the human brain: they will not develop as a child does. From birth onward, the cortex forms as a result of the proliferation, migration and pruning of neurons and of a process we call plasticity that is highly dependent on experience. Our models will instead begin at any arbitrary age, leapfrogging years of development, and

---

[4] Markram, "A Countdown to a Digital Simulation of Every Last Neuron in the Human Brain," p. 52.
[5] Ibid.
[6] The simulation of central processes of biology, like protein folding for example, seems to pose extreme difficulties computationally, so simulating life on the molecular level might only be possible with models into which a lot of detailed information e.g. on the rates of certain chemical reactions, is preprogrammed in advance. Such models cannot be used to predict novel phenomena not yet foreseen by their designers, e.g. on the effects of drugs.
[7] I think it is not – and a lot could be said about that - and I think the article is basically a piece of propaganda from a time when he was trying to get funding and support for his project.

> continue from there to capture experiences. We will need to build the machinery to allow the model to change in response to input from the environment.
>
> The litmus test of the virtual brain will come when we connect it up to a virtual software representation of a body and place it in a realistic virtual environment. Then the in silico brain will be capable of receiving information from the environment and acting on it. Only after this achievement will we be able to teach it skills and judge if it is truly intelligent. Because we know there is redundancy in the human brain—that is, one neural system can compensate for another—we can begin to find which aspects of brain function are essential to intelligent behavior.[8]

Correspondingly, I think that such a project raises some philosophical, especially ethical, problems. So in the rest of this essay, I am going to concentrate on these issues, although by no means in an exhaustive way.[9]

III. Ethical Issues

The prospect of the possibility of the simulation of a whole human brain, as described in Markram's article, raises the question: could such a simulated brain become conscious? Indeed, in a 2013 interview in *New Scientist*, Markram was directly asked precisely that question: "Once completed, could the simulated brain ever become conscious?" He responded:

---

[8] Ibid., p. 55.

[9] The argumentation presented below is based on certain ideas in the philosophy of mind. Basically, my view is a non-reductive version of computational functionalism. It would definitely be interesting to go deeper into issues of the philosophy of mind here, but due to current limitations of my time caused by private matters, I am not able to discuss these interesting aspects in more detail the context of this paper. Let me just briefly outline some aspects of my position: I think consciousness is real, so I think eliminative as well as behaviorist approaches to the problem are wrong. I am not a dualist and in the traditional debate between idealism and materialism (or physicalism), I would opt for the materialist or physicalist side. As I mentioned, I would characterize my position as a form of computational functionalism, although my concept of "computational" goes beyond Turing-computability and includes creative processes that cannot be formalized in their entirety; see, e.g., A. Keller, "Proteons: Towards a Philosophy of Creativity," *Borderless Philosophy* 2 (2019): 117 – 172, available online at URL = <https://www.cckp.space/single-post/2019/06/01/BP2-2019-Proteons-Towards-a-Philosophy-of-Creativity-pp-117-172)>. Briefly put, I think that consciousness flows naturally and necessarily, but *not* superveniently, from the creative information-processing structure of our nervous systems. So I think that consciousness is not an epiphenomenon, and also that there is no metaphysical possibility of "zombies"—see, e.g., R. Kirk, "Zombies," *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), E.N. Zalta (ed.), available online at URL = <https://plato.stanford.edu/archives/spr2019/entries/zombies/)>. A few more hints about the direction my thoughts are taking concerning the mind-body relation can be found towards the end of this essay.

> When we couple the model to a robot, the robot will behave, and we'll see this in the way its neurons are firing. Does that mean it's conscious? That's a philosophical question—and an unresolved one. [10]

This answer, I think, shows that Markram was quite uninterested in the question. Perhaps he found the issue cumbersome. If you take it seriously, you must also take the ethical questions it entails seriously. Doing so might have meant stopping the project, at least as originally conceived.

Markram's original plan, as presented in the 2012 *Scientific American* article, turned out to be overambitious and unfeasible within the time and resources of the project; so the project was eventually reorganized, with more modest goals. But here I am taking the scenario presented in that article as a *thought experiment*. What Markram proposed *might* become technologically feasible, although it might also be prohibitively expensive.

Why is there ethics at all? Why is nothing wrong with taking a hammer and beating a stone, but everything wrong with taking a hammer and beating a human being? The reason, I think, is simple: the human being would suffer, the stone would not. We assume that animals, at least the more sophisticated ones, also have such a subjective side, so we would extend some of the ethical principles to them as well. There is obviously no point in giving a dog the right of free speech or free choice of religion, but we should not beat it with a hammer. The subjective experience of the human being or animal, especially the possibility to experience suffering or enjoyment, is what makes ethics necessary.

Being "artificial" vs. being "natural" cannot be the relevant criterion here. The distinction is a doubtful one anyway. If I am standing near a river in a big city, looking at the riverbank, am I seeing something artificial or something natura? In many places, the riverbank consist of walls or banked-up stones, so to that extent, the river is a technological product and structure. In the age of genetic technology, building a living organism from scratch is getting within reach. Every artificial system has natural components, and every natural system operates according to the same laws of physics. So, the natural/artificial distinction is not as fundamental as it might seem. But if a system we built artificially were sentient, i.e., were able to suffer or enjoy, then that alone would qualify it for ethical consideration and treatment.

If it is physically or even technologically possible to build a machine in such a way that it is sentient, and thereby has subjective experience, we must apply rules of ethics to it. This would apply to an "artificial animal," even if it is not sapient or

[10] J. Griggs, "Why We're Building a €1 Billion Model of a Human Brain," *New Scientist* (6 February 2013), available online at URL = <https://www.newscientist.com/article/mg21729036-800-why-were-building-a-1-billion-model-of-a-human-brain/>.

intelligent. If it were sapient or intelligent, then we would have to grant it certain rights. It might be a form of intelligence that is totally unlike our own. An artificial intelligent sentient system might not have a human body, and might not have human-like experiences, but it might also have a form of experience that is rationally like our own and therefore would require giving it some rights.

At the time I am writing this, we are technologically still a long way from this vision. The state of the art is to simulate something like a brain column of about one square millimeter, i.e., something on the "circuits" level in Markram's hierarchy, but this simulation requires a roomful of hardware, and does not even go down to the molecular level.[11] Nevertheless, although Markram's estimates were obviously over-optimistic, we cannot be sure that a whole-brain simulation would be physically or even technologically impossible.

Now let us consider the possibility—discussed below in more detail—that a human-like consciousness indeed arises in such a simulation. If the simulated brain were to contain a human-like conscious mind capable of experiencing emotions, pain and so on, then that conscious mind would have to be regarded as a *human being*, in the moral sense (as opposed to the merely biological sense), that he or she or they would have to be granted full human rights, including (i) the right to live (so s/he or they must not be switched off and deleted), (ii) the right of physical integrity (so s/he or they must not be used for experiments on the effect of, say, simulated brain injury or simulated brain diseases or drugs and so on), (iii) the right not to be tortured (so s/he or they must not intentionally caused to experience an excessive level of pain or suffering), (iv) the right of freedom (so using him or her in any way would have to be regarded as slavery), (v) the right to privacy (so we must not spy inside his or her mind), and so on. If we regard it as immoral to put somebody into a state of hibernation without her or his consent (assuming this was technologically feasible) for prolonged times, perhaps forever, then we would also have to regard it as immoral to switch off the simulated brain even if all the data were preserved in such a way that it would be theoretically possible to switch it on again.

In fact, if the simulated brain contains a human consciousness, then the mere act of creating it would be immoral. We would have, in effect, created a newborn baby with Locked-In-Syndrome, with the full emotional and cognitive needs of a baby and yet no way to satisfy these needs. That would be a cruel thing to do. One could argue that the development of emotions (i.e., desires, feelings, and passions) heavily depends on the environment (both the body and the outer environment) in which an infant brain normally develops. But we cannot simply assume that in the absence of such an environment, there would not be at least some minimal and

[11] See E. Gent, "Supercomputer Simulates 77,000 Neurons in the Brain in Real-Time," *New Scientist* (8 October 2019), available online at URL = <https://www.newscientist.com/article/2218993-supercomputer-simulates-77000-neurons-in-the-brain-in-real-time/>.

potentially suffering consciousness. We would have created a severely disabled but still sentient entity.[12]

Markram correctly writes that "ethical concerns restrict the experiments that neuroscientists can perform on the human brain." If I am right, the same restrictions would hold for a *simulated* human brain. A*ll ethical constraints that apply to experiments on human beings must apply here too*. The ability to experience joy and suffering is the reason for treating humans not merely as things. It is the basis of ethics. We do not understand yet how subjective experiences of joy or pain can arise in our conscious minds, but we should assume that they also would arise in a simulated brain, unless we can prove otherwise.

We should consider the resulting philosophical and legal questions before we ever consider doing such an "experiment." We are stirring up a hornet's nest of philosophical and legal problems here. For example, if you connect the simulated conscious brain to an artificial body, the resulting robotic system would have to be regarded as a human being in the moral sense, with all the rights of a human being, including a right to live, to be free, to have privacy, to decide over his or her own life and so on. If that artificial human being decides to become a musician, a Buddhist monk, or a carpenter, we will have to respect his or her decision. We will have to continue paying the bill for those super computers, because switching them off would be murder. If he or she decides to end his/her/their life, then if we think humans have a right to commit suicide, we will have to grant him/her/them the right to destroy their own artificial brain. What if s/he or they becomes a criminal? I leave it to the reader to think such questions through in more depth.

The possibility of artificial entities with subjective experience requires ethics to be put on a foundation that is not dependent on the notions of human dignity, human freedom,[13] or the assumption of something like a "transcendental self." Instead, it would be based solely on the fact that subjective experiences can be either pleasant or (especially) unpleasant or even painful to the point of suffereing, and that we assume that others have such experiences as well. Additionally, we would have to subscribe to some principle of empathy, i.e., we must not be indifferent to the suffering of others. Not assuming such a principle as an additional "axiom" would lead to a quasi-psychopathic doctrine of egoism, and I think that everybody has the good reason to resist such a doctrine. Concepts like "dignity" might be derivable

[12] When using the term "consciousness" in this article, I mean this primarily in the sense of *phenomenal* consciousness, i.e., subjective experience, and not primarily in the sense of a self-conscious awareness that requires a more advanced cognitive development than the initial simulated brain would be likely to have.

[13] One might indeed ask the question whether libertarian freedom in the classical agent-causal sense is even metaphysically possible, whether in a simulated brain or a biological one. I actually have doubts about this, but that is beyond the scope of this article.

inside such an ethics, but it would also apply to non-human animals and to other sentient entities, even artificial ones.

Obviously I have not fully worked out this approach and I do not expect this to be easy, but the possibility of artificial experiencing entities seems to require that we do some serious thinking about these issues. Insisting that experiencing machines are impossible is, in my view, not an option unless we can prove it. In the interview, Markram said that it was an open question. In my view, as long as this question is open, i.e., as long as we cannot exclude the possibility of pleasant or painful subjective experience in such machines, the consequence must be that we must not build them.

In the same connection, here is one more thought: if human-like consciousness could be "implemented" by a machine, would that devalue human beings or their subjective experience? I think that this is not the case. Saying something like "humans 'are only' machines," "love 'is only' a brain process," is, in my view, a mistake. Using that devaluating term "only" is, I think, a wrong-headed pattern of thought. It is based on taking a false point of view. From the point of view of a scientist, the process of smelling a rose might indeed just appear to be a chemical process. From the point of view of the conscious being (what I call the "internal observer" below), it is something categorically different. That the scientific point of view is possible, does not mean it is the one we must take all the time or that it invalidates the subjective point of view. Instead, the subjective point of view is primary and the source of all values.

From the point of view of the universe, our planet is just like a dust grain. From the subjective point of view, it is the world, and the universe is an abstract entity without much relevance. Nothing forces us to take on the point of view of the universe if we are outside the point of view of astronomy or cosmology, say. Values emerge from the fact of subjective experience and the subjective point of view is primary. The nature of the underlying entities does not matter. From the subjective point of view, the question whether the subjective experience is a result of physical processes or of some special non-physical mental substance simply does not matter. It is interesting only from the point of view of science and philosophy, i.e., from an outside point of view that we normally do not take and that we are not forced to take. Subjective experience is immediate, and to that extent, its underlying (meta)physical nature is irrelevant from its own point of view, however interesting that question might be theoretically.

IV. Consciousness in the Machine?

In section III, I touched upon some of the ethical problems that would result if a whole brain simulation had consciousness. In this section, I'll argue that this might

indeed be the case, and that the subjective qualities of such an "in-silico consciousness" would be like those of a human being.

I'll starting from the first-person fact of consciousness in human beings. Introspectively, I am aware of myself as conscious, with sensations and emotions, including at times pain or other forms of suffering. Moreover, I can think, talk, and write about this experience, so in some way it influences explicit, propositional thinking. This means, I think, that it is *not* a mere epiphenomenon, caused by something else but without any causal powers of its own. I do not know how it comes about, but since I am not a solipsist, I also believe that every human being in the moral sense also has subjective experience. I therefore hold that causally efficacious human consciousness somehow emerges from physical processes and structures in the human body and especially its nervous system.

Now, if we introspectively investigate our own consciousness, we will find that we are aware neither of neuronal or molecular processes, nor of most of the information processing underlying our cognitive and perceptive processes. The internal details of these processes are not introspectively accessible. The movements of blood cells through the brain's capillaries, the movements of vesicles at the synapses, the exchange of molecules at the membranes of neurons are not consciously accessible to us. Likewise, much neuronal information processing obviously happens outside of our consciousness. For example, much of the detail of controlling movements of our body appears to be "automatic."[14] Some processes and entities are accessible to our introspection, others are not. There is a "*horizon of accessibility*" that we cannot see through "from the inside."

Let us assume, as a thought experiment, that a whole brain simulation can be built. This simulation contains simulated neurons. Simulated nerve impulses are traveling through those neurons. Simulated vesicles at simulated synapses discharge simulated neurotransmitters.

Now let us also assume a natural, human brain having a certain subjective experience, for example, smelling a rose. This activity in the olfactory bulb creates a certain pattern of neural impulses going through the brain. If consciousness emerges from the structure of this process, in whatever way, as we have assumed, then this pattern of neural impulses must somehow generate the subjective experience of the scent of a rose. Now, if we cause the corresponding, simulated pattern of nerve impulses in the simulated brain, then would there be the same subjective experience? The neurons and the processes inside them are not perceptible to our subjective experience. Molecules in them are constantly exchanged with others, as blood is

[14] One could use the term "subconscious" here, but I would like to avoid the burden of ideas and connotations with which this term is associated in psychoanalytic traditions, such as Freud's and Jung's.

flowing through the brain and exchanges molecules with it. The microscopic properties of the neurons are constantly changing. These changes, as far as they are part of the normal brain metabolism, do not affect subjective experience. Certain components may be exchanged, e.g., you may receive a unit of blood from a blood donor, so there suddenly are blood cells with different genetic sequences in your brain. So, there are changes that can happen in the molecular structure that do not affect the conscious being although the conscious being is implemented in terms of these processes. There is a "horizon of accessibility." Events behind that horizon are not accessible to the conscious being. They happen in its "automatic infrastructure"[15].

Such horizons also exist in technological systems. If you publish an article on a blogging platform, for example, then your data might be moved from a server in one room or even one city to another server in another place, in processes of "load balancing" or "failover," for example, or from a hard disk to flash memory or an optical disk. Some properties of the system are hidden. They are behind a horizon of accessibility from the point of view of the user, or from the point of view of an application.

A simple example might clarify this point: consider a text file stored on a hard disk. Now copy it to a USB stick or a CD-ROM. The physical representation of the file will be completely different in each case, but any application using the file cannot "see" these differences. The operating system and device drivers create a world of "emulated objects" whose properties can be described and understood independently of the physical system used to implement them. On a physical level of description, what you have are magnetic orientations of particles, small holes in the CD-ROM's surface, electron distributions and so on, together with processes "reading" these features that are physically completely different from each other.

[15] Another term used in computer science for this hiddenness of infrastructure or information is "transparency." The term was first used in the context of computer networks. If a message entering a computer network on one side comes out on the other side unchanged, the network is transparent for that message. During the process of transmitting the message, it may be compressed, encrypted, chopped into packets that are sent along different lines, reassembled, decrypted and decompressed, e.g., but it is not possible to infer any of these processes from the message that comes out. The metaphor of transparency here was used in analogy to a light signal going through a transparent medium, like a pane of glass. So, in this use of the term "transparency," it means that we do *not* see the details. So "transparency" in this sense means *invisibility*. Obviously this can be sometimes confusing, since there are other uses of the word that mean just the opposite. I have not traced the origin of this term, I first encountered it in the 1980s, and it might be one or two decades older. The term has since been expanded to other cases in which the details of an infrastructure are inaccessible to an application, especially in the context of "cloud computing," where we can, for example, say that the servers running a program are transparent to the program or to the user (so you might use a "serverless" service although actually there is some hardware in each instance). As an antonym to "transparency", the term "awareness" is sometimes used. Note that in the term "cloud computing" the inaccessibility of something is represented by the opacity of a cloud, so, paradoxically, the very same phenomenon of information-hiding is being expressed via the opposing metaphors of transparency and opacity.

From the application's point of view, these differences are not accessible, and the usefulness of computers to a great extent comes from the possibility of creating such emulated objects and "everting" them to us through a user interface. The application resides in a world of objects that are, in a way, independent of the underlying physics. It is an observer of emulated objects.

You might say that this emulated world is just a layer of description and that "in reality" the only thing existing is the magnetized particles, electrical currents, and so-on, and that the application which we as programmers are thinking about exists only as a description in our minds, while in reality, there is a physical machine executing machine instructions only. But a crucial aspect of consciousness seems to be its self-awareness or immanent reflexivity. If the "application" can observe itself and is itself part of the world of emulated objects, then being able to create descriptions of itself and of the processes inside itself, it will be existing from its own point of view, thus acquiring consciousness. We would then have, inside the system, an "internal observer" that is emulated by the system and that exists from its own point of view. And this observer would have a horizon of accessibility shielding the details of its implementation from its own point of view.[16]

My idea is that our own consciousness is an immanently reflexive internal observer of this kind, emulated by the neuronal processes in our brains. If you simulate these processes in a computer system, such an internal observer would be present too. The brain would be *simulated*, but the consciousness inside would be *emulated*, *just like the one in a biological brain*. The details of its hardware (biological vs. silicon) would be *beyond its horizon of accessibility*. This means that the resulting conscious mind would, in a subjectively experiential way, be just the same as the conscious mind of a biological brain having the simulated structure. The nature of the "hardware" implementing the conscious mind is *beyond* the horizon, so the mind cannot distinguish whether it runs on natural or simulated neurons. Therefore, if there is a consciousness present, then the nature of that "hardware" does not matter for the question of the ontological or moral status of the resulting consciousness. A conscious silicon-based mind therefore would have to be treated just like a conscious biological one.

Here is another point. Achieving consciousness inside a simulated brain might be computationally easier than doing a whole brain simulation down to the molecular or neural level, because if we replace detailed simulations of some of the brain's components (e.g., neuronal columns) with simplified approximations taking less computational resources, those details could again be beyond the horizon of

[16] I don't mean this in the sense of a little homunculus sitting in the mind (with another homunculus sitting in the mind of the homunculus, etc.), but instead in the sense of an information processing process that exists from its own point of view and for which the emulated word made accessible to it by the underlying processes is the primary reality.

accessibility of an internal observer arising within the system. So. we might get into the "danger zone" of creating a human conscious mind even before we reach the technological ability to simulate a human brain down to the neuronal or molecular level. I therefore suggest thinking those philosophical questions through very carefully before such experiments are undertaken.

I am not going to go much deeper into the philosophy of mind in this context. I do, however, glimpse the possibility of a defensible approach here. I think that the idea of transparency = the invisibility of certain kinds of information, or the "horizon of accessibility," provides a way to think about layers of objects that exist in a quasi-independent way, although the "higher" layers are implemented in terms of the "lower" ones. Consciousness in the brain may exist in an ontologically irreducible way precisely because of the existence of this horizon. The mechanical components pushing each other around, to use Leibniz's metaphor of the mill,[17] are invisible from the point of view of that entity. It is emulated by the biological or technological system. Such an emulated system can be described and understood independently of the physical properties of the emulator it is running on. If it is an information processing system, a system that can be described as an observer, it might have a perception of itself, i.e. it might exist from its own point of view. It is not identical to the system that emulates it, because we could exchange components in the underlying system without affecting the emulated system. So, there is an "ontological transition," from a world of neurons to an emulated, but nonetheless real, world of conscious phenomena, a world that exists from its own point of view and thus has a degree of ontological autonomy.

Therefore, in this approach to mentalistic ontology, ontologically autonomous objects are created by hiding some aspects of their implementation. From the point of view of the underlying objects, the emulated object does not exist. From the point of view of the emulated object, the hidden objects of the implementation do, in some sense, not exist as well. This is not reductionism, because the system can be described without referring to the underlying system, although it is also implemented in terms of the underlying system. I do not claim that this is the whole story, for the "hard problem" partially remains, but I do see a promising approach here that is different from standard versions of non-reductive computational functionalism that rely on natural or nomological supervenience, and thereby entail epiphenomenalism. The

[17] Leibniz, in section 17 of the *Monadology*, wrote about the question of sentient machines: "It must be confessed, moreover, that perception, and that which depends on it, are inexplicable by mechanical causes, that is, by figures and motions, And, supposing that there were a mechanism so constructed as to think, feel and have perception, we might enter it as into a mill. And this granted, we should only find on visiting it, pieces which push one against another, but never anything by which to explain a perception. This must be sought, therefore, in the simple substance, and not in the composite or in the machine." See G.W. Leibniz, *The Monadology*, trans. R. Latta (1898, originally published in 1714), available online at URL = <https://www.plato-philosophy.org/wp-content/uploads/2016/07/The-Monadology-1714-by-Gottfried-Wilhelm-LEIBNIZ-1646-1716.pdf>.

experiences are not, as the essentially dualistic zombie-theory assumes, something that could just as well not be there. As indicated above, we are immanently reflexively aware of our own subjective experiences, so in order to understand how they are possible, we would have to study how such immanently reflexive awareness can come about inside an experiencing system, and such studies might open up the way to solving the problem.

V. Conclusion

Projects like The Human Brain Project might bring us a step further to understanding these issues and thereby turn the philosophy of mind into a natural science. But I believe that we should actually stop the simulations before they cross the threshold of becoming conscious, in the sense of being able to subjectively experience any kind of suffering. In terms of the "boxological" graphics in Markram's article,[18] I believe that for good ethical reasons we should pull back before we reach the "regions" level, or start going beyond it.

This conclusion in turn yields a perhaps surprising implication. Finding out how consciousness really works might necessarily involve observing mental processes in a simulated brain. But building such a simulation would arguably be immoral. If so, then a complete solution for the "hard problem" would be blocked by ethical considerations. And were it to be that way, then let it be so.

[18] Markram, "A Countdown to a Digital Simulation of Every Last Neuron in the Human Brain," p. 38.

References

Gent, E., "Supercomputer Simulates 77,000 Neurons in the Brain in Real-Time." *New Scientist* (8 October 2019). Available online at URL = <https://www.newscientist.com/article/2218993-supercomputer-simulates-77000-neurons-in-the-brain-in-real-time/>.

Griggs, J., "Why We're Building a €1 Billion Model of a Human Brain." *New Scientist* (6 February 2013). Available online at URL = <https://www.newscientist.com/article/mg21729036-800-why-were-building-a-1-billion-model-of-a-human-brain/>.

Keller, A., "Proteons: Towards a Philosophy of Creativity," *Borderless Philosophy* 2 (2019): 117 – 172. Available online at URL = <https://www.cckp.space/single-post/2019/06/01/BP2-2019-Proteons-Towards-a-Philosophy-of-Creativity-pp-117-172)>.

Kirk, R., "Zombies." *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), ed. E.N. Zalta. Available online at URL = <https://plato.stanford.edu/archives/spr2019/entries/zombies/)>.

Leibniz, G.W., *The Monadology*. Trans. R. Latta (1898, originally published in 1714). Available online at URL = <https://www.plato-philosophy.org/wp-content/uploads/2016/07/The-Monadology-1714-by-Gottfried-Wilhelm-LEIBNIZ-1646-1716.pdf>.

Markram, H., "A Countdown to a Digital Simulation of Every Last Neuron in the Human Brain." *Scientific American* 306 (2012): 34-39. Also available online at URL = <https://www.scientificamerican.com/article/human-brain-project-digital-simulation-neuron/>.

Theil, S., "Why the Human Brain Project Went Wrong—and How to Fix It." *Scientific American* 313 (2015): 36-42. Also available online at URL = <<https://www.scientificamerican.com/article/why-the-human-brain-project-went-wrong-and-how-to-fix-it/>.