Is Consciousness Supervenient on the Physical?

Jaemoon Lee

Introduction

John McCarthy, a pioneer of artificial intelligence, stated that the following about the goal of artificial intelligence as a field [6]:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

Since the beginning of AI, developments have been made to simulate a variety of features of intelligence such as image classification, and path planning. In the natural world, intelligence is found alongside consciousness. It is therefore a natural question to ask: Can machines also be made to simulate consciousness?

As stated by McCarthy, a first step for investigating the possibility of consciousness in machines is to understand consciousness precisely. A first question may be the following: Can consciousness even be explained in terms of the physical? In contrast to many features of intelligence, it is difficult to say with certainty that consciousness is caused by the physical. In fact, the philosophy community has long been grappling with this question [1, 2, 4, 7]. Supervenience is a philosophical concept that formalizes the idea that some set of properties of an entity depend upon another set of properties [4]. In this essay, I consider whether the state of consciousness of an entity supervenes upon its physical properties.

Supervenience

In this section, I will discuss the philosophical concept of supervenience. Supervenience is defined as follows:

Definition 1 *Supervenience [2] Let A and B be two sets of properties. A supervenes upon B if no two possible situations can differ with respect to their A properties without also differing with respect to their B properties.*

This is logically equivalent to saying that if the B properties are the same, then the A properties must be the same. Supervenience formalizes the idea that the A properties depend upon the B properties. In order to illustrate the concept of supervenience, I will present a few examples.

Example 1: The area and radius of a circle Consider circles. Let A be the area of a circle, and B the radius of a circle. If two circles have the same radii, then they have the same area since the area of a circle is determined by its radius via the equation $A = \pi r^2$. Therefore A supervenes upon B.

Example 2: Biological and physical facts [2] Consider all possible worlds. Let A be the set of all biological facts, and B the set of all physical facts. For example, how the process of DNA replication occurs in a world would belong in A, while how gravity works in a world would belong in B. If two worlds have the same physics facts, then the biological facts must be the same. Therefore A supervenes upon B.

Example 3: The pressure of gas [2] Consider a single mole of gas. Let the A properties be the pressure of the gas, and the B properties the volume and temperature of the gas. The ideal gas law states that the pressure P, volume V, and temperature T have the relationship that $PV = KT$ where K is a constant. Therefore the pressure of the gas is supervenient upon its volume and temperature.

Local and global supervenience

Supervenience may be categorized into local or global supervenience [2]. Whether supervenience is local or global depends on what the 'situations' are in the definition of supervenience. If the situation with the properties in question is an individual member of a world, then I am considering local supervenience. In contrast, if the situation with the properties in question *is a world*, then I am considering global supervenience. For example, the supervenience described in Examples 1 and 3 were local, while the supervenience described in Example 2 was global. In this essay, the supervenience in question is local supervenience.

Logical and natural supervenience

A second categorization of supervenience is into logical and natural supervenience [2]. Natural supervenience means that the B properties will necessarily enforce the A properties of entities in the same world. Logical supervenience means that the B properties will necessarily enforce the A properties of entities even in different worlds. An alternative description of the difference between logical and natural supervenience is that it depends upon what I mean by 'possible' in the definition of supervenience. A is logically supervenient on B if A is supervenient on B in any conceptually coherent world. One trivial example is that A logically supervenes on itself. In contrast, if A is supervenient on B within the context of a certain world or classes of worlds, but it is conceivable that there be a world where A is not supervenient on B, then A is naturally supervenient on B. For example, biological laws directly follow from physical laws and therefore Example 2 is logical supervenience [2]. On the other hand, one may argue that it is conceivable that in some universe pressure does not depend the same way on volume and temperature [2]. By a similar argument, one may say that it is conceivable that in some universe the area of the circle depends differently upon the radius. Therefore Examples 1 and 3 are natural supervenience. Logical and natural supervenience have also been referred to as 'strong' and 'weak' respectively [7].

The supervenience of consciousness on the physical

In this section, I will present a few perspectives on the question of whether consciousness is supervenient on the physical. In other words, if two entities are exactly the same physically, can one be conscious and the other not? Or can their consciousness deviate in any way? Before presenting the perspectives, I will discuss what consciousness is.

Understanding consciousness

Much of the difficulty with analyzing the supervenience of consciousness on the physical is precisely defining and explaining consciousness. Some properties associated with consciousness can be explained scientifically: the ability of an entity to react to its environment, to absorb information from its environment and incorporate the information into its own mental system, and to deliberately control its behavior [1]. But in addition to these properties, there is a notion of experience for consciousness that is more difficult to pin down and explain, i.e. to be conscious is to experience myself and my environment [1].

Some philosophers have argued that consciousness is best explained by viewing it as an *emergent process* [3, 8]. An emergent process is a novel process that arises on top of some other processes. Because the emphasis with an emergent process is that it be more than the sum of its parts, it is considered distinct from supervenience [3]. Viewing consciousness as an emergent process is backed up by neuroscience [11]: It is well-known that cognitive processes are not local processes in the brain but rather require complex interactions involving many areas. Therefore, it has been proposed to not explain consciousness at the neural level but rather look to dynamical patterns of activity going on in the brain over time. Consciousness is the integration and combination of all these individual cognitive processes, and therefore an emergent process of the brain.

For practical purposes of considering the consciousness of machines, a scientific understanding of consciousness is needed. It has been argued that such an understanding is possible [10]. One possible direction is by associating consciousness with patterns in the brain that appear to correlate with these states [10]. Despite this theory, it is still an open problem what relationship consciousness has with the physical brain [5].

Is consciousness logically supervenient on the physical?

I first considered whether consciousness supervenes logically on the physical. Because I am considering logical supervenience, this means whether or not it is conceivable that some world exist where entities can be physically identical, but have different states of consciousness.

A commonly used argument for the perspective that consciousness is not logically supervenient on the physical rests upon the difficulty of disproving that lack of consciousness is inconceivable [2]. Consider the 'zombie' argument [2]: Suppose that in some world, I have a physically identical twin. The twin processes information the same that I do, reacts to stimuli the same that I do, moves the same that I do, etc.. However, the twin is not conscious at all, hence a zombie. Is this conceivable? One might imagine that perhaps the zombie has been conscious for a different length of time, and consciousness evolves and changes as a function of time in a way not determined by the physical state of the brain. It is difficult to refute the conceivability of this situation: In order to argue that the zombie is inconceivable, one must present a compelling argument about where exactly this situation doesn't make sense.

The fact that there does not currently exist a definition for consciousness in objective, measurable terms contributes to the difficulty of arguing against the previously described position. For contrast, consider the question of whether memory is logically supervenient on the physical. Memory is defined as how the mind stores and remembers information, and there exist scientific explanations for memory [9]. Because of this direct link with the physical, it is much less conceivable that a physically identical twin have different memory.

Arguments against the logical supervenience of consciousness upon the physical are not completely satisfying. The reason is that because they rest upon the difficulty of coming up with a counter example. But, is this difficulty because there is some fundamental truth in the argument, or is it because I do not have the necessary definitions and scientific facts to convincingly disprove the argument? Consider again the supervenience of the biological facts in a world upon the physical facts in a world described in Example 2, which has been regarded as an example of logical supervenience [2]. Suppose one were to ask a person with no knowledge of modern science to describe why it is inconceivable that two worlds have different underlying physical facts but different biological ones. Then, they probably would not have a counter example as they have no knowledge of how the physical facts would cause the biological ones. With this example in mind, do I know that consciousness could not be described in a concrete way that would make more clear how I could go about testing its relationship with the physical? Do I know that there could not be future scientific discoveries that further make this link as clear as that between the biological and the physical? Methods of getting around this argument could be restricting my definition of 'the physical' to only include those facts currently known.

Is consciousness naturally supervenient on the physical?

On the other hand, I considered whether consciousness is *naturally* supervenient on the physical. Meaning, in the world that we live in, can two entities have different states of consciousness while being the same physically? I believe this is a more relevant question when considering the consciousness of machines. After all, machines are created in our world and not in some philosophical alternative.

Perhaps there is not enough scientific evidence to answer this question conclusively, but my intuition is that consciousness is naturally supervenient on the physical. My intuition is motivated by considering the effects that many physical events can have on my consciousness. Consider falling asleep, which manifests as physical changes in the body that can be detected by means such as a heart rate monitor. Sleeping at least appears to change the conscious state, even appearing to have temporarily ceased from the perspective of the individual upon waking up. It is certainly believable that some changes in the body cause the state of sleep. Alternatively, consider a drug or illness that alters the conscious state of an individual. It is again believable that physical changes from the drug or illness are bringing upon the change in conscious state.

What does this imply for AI?

Now that I have considered whether consciousness is supervenient upon the physical, I returned to my initial consideration: Can machines simulate consciousness? As I argued in the previous section, natural supervenience is in fact more relevant to answering this question, since I am building machines in our world. If consciousness is naturally supervenient on the physical, then if a machine is physically the same as a conscious being in those key physical properties, then it should be in the same state of consciousness. It is difficult to determine whether a machine could exactly imitate

those key physical properties without knowing what they were. But, I don't find any clear objection to the possibility that any part of the brain could be simulated by a machine.

However, considering whether it is possible to simulate consciousness is just the beginning of the problem. Perhaps the most difficult hurdle would be to understand consciousness on a scientific level enough to reproduce the physical conditions on a machine. But beyond that there would be ethical considerations: Is it right to create consciousness? If a machine is conscious, is it feeling pain and discomfort? It is also important to consider that consciousness may not be something that I purposefully create: perhaps I could accidentally create machines are conscious, and must investigate the implications for that.

Conclusions

In this essay, I have considered whether consciousness is supervenient upon the physical. My motivation for this problem is to understand consciousness and its physical causes in order to determine whether machines could be made to be conscious by imitating conscious beings physically. Consciousness is at least not logically supervenient on the physical, but whether it is naturally supervenient is an open problem. A greater scientific understanding of consciousness and its relationship with the physical is most likely needed in order to better address this problem.

References

[1] D. J. Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

[2] D. J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford university press, 1996.

[3] P. Humphreys. Emergence, not supervenience. *Philosophy of science*, 64:S337–S345, 1997.

[4] J. Kim. *Supervenience and mind: Selected philosophical essays*. Cambridge University Press, 1993.

[5] C. Koch. What is consciousness?, 2018.

[6] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12, 2006.

[7] B. McLaughlin and K. Bennett. Supervenience. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018.

[8] B. P. McLaughlin. Emergence and supervenience. *Intellectica*, 25(2):25–43, 1997.

[9] N. News. How does memory work?, 2016.

[10] U. T. Place. Is consciousness a brain process? In *The mind-brain identity theory*, pages 42–51. Springer, 1970.

[11] E. Thompson and F. J. Varela. Radical embodiment: neural dynamics and consciousness. *Trends in cognitive sciences*, 5(10):418–425, 2001.