Can mathematical structures be conscious?

Brains are conscious because of the computations that they perform, and if a computer is capable of simulating a brain (and the world it interacts with), it would therefore be conscious as well. Max Tegmark goes one step further in suggesting that not even a computer is needed to create a conscious being: a computer simulation can be represented as a static four-dimensional object, and this object arguably exists as a mathematical structure even if the computer were to disappear altogether. By this argument, there are mathematical structures describing computer simulations contain conscious entities, and feel as real to their inhabitants as simulated universes or 'real' universes such as our own. If this is true, then there are a vast number of mathematically possible universes with the same claim to physical existence as our own; and the existence of our universe becomes indistinguishable from the existence of the mathematical structure that describes our universe, and hence our universe is effectively just a mathematical structure.

This essay focuses on the critical part of Tegmark's argument: can mathematical objects, as opposed to computer simulations, be conscious? It is not an exaggeration to call this question potentially the one of the most important ever asked, as the physical existence of countless mathematically possible universes and the entities that they contain hangs on the answer. If the answer to the title question is yes, then it would also be the answer to fundamental questions such as whether our universe had a beginning and why it is here; our universe would be a mathematical object, whose existence is as timeless and inevitable as the existence of the number four. It would give definitions of otherwise elusive concepts such as existence and consciousness, which turn out to be straightforwardly mathematical notions. Physics would be reduced in turn to a branch of mathematics, that branch dealing with formal systems capable of generating self-replicating things that can model the world around them and develop consciousness.

Because of the far-reaching implications of Tegmark's argument, it is important to ask whether it is in fact correct. What follows is a review of the critical part of Tegmark's account, and then some possible arguments against it. In particular, despite the enormous empirical difference between the two potential answers, it turns out to be remarkably difficult to make a version of the main question 'Are mathematical objects conscious?' that is testable. Despite this problem of testability and some other possible holes in the argument, there is one part of it that is undeniable, which is that formal systems can describe (and even be, at least while they are being calculated) universes as complex and worth exploring as our own. We are in principle able to calculate these formal systems that describe universes, and to look into the mental states of their inhabitants, making the study of our own universe merely a tiny corner of a vast subject, the study of formal systems. Mathematical objects can be identical to conscious beings in all of their essential details, even if they lack some final spark of subjective consciousness that comes from being simulated. The essay concludes that the profound implications of this fact transcend the possibly unknowable question of whether entities in these mathematical objects are really conscious or not.

To review Tegmark's argument as described in Tegmark (2007) and Tegmark (2014:347-350), one normally can accept that a brain can be simulated on a computer, with a toy world to interact with, and this would feel as real to the simulated brain as our world feels to us. This is because the human brain is indistinguishable in the operations that it is performing from what a computer is able to do, namely to process information. The next step is to imagine a computer not simulating it, but simply describing this toy world, which

might mean writing out the output of the simulation and having some description of the relationships between each time-slice. Alternatively, the program itself could be the description of this toy world, as when any computer runs the program, then the same toy world will be produced each time. This data can then be stored in some form, on a computer or even on a USB stick. Tegmark then says 'It would appear absurd that the existence of this memory stick would have any impact whatsoever on whether the universe it describes exists "for real"' (Tegmark 2014:349.

There are two possible ways of dissenting from this argument. First, it is possible that a computer carrying out a simulation may not be equivalent to a computer storing four-dimensional data. The important property of a computer is that it takes inputs and produce a previously unknown output. When a computer is carrying out a calculation, one can sense that something is happening which calls a universe into existence. This is not clear for a static representation of the simulation, which could be stored on a piece of paper. Unlike a computer carrying out a simulation, the piece of paper lacks the ability to demonstrate to us the details of the universe that the information describes.

A second way of dissenting from Tegmark's argument is to assert that a physical object needs to exist to be representing the structure for it to be conscious. Even if it appears absurd that the existence of a memory stick is needed for a universe to exist, its intuitive absurdity is not sufficient to rule this possibility out. A different version of this point is that Tegmark may be confusing two different readings of mathematical statements. The statement 'triangles have three sides' is a convenient shorthand for saying that if a physical object is triangular, then it will have three sides. This does not commit to the existence of triangular objects in the world. Similarly, 'unicorns are white' can be true as a definitional property of unicorns, but does not commit to the existence of unicorns. This statement is therefore true in a hypothetical sense (if there was a unicorn it would be white) but not in an actual sense. Perhaps the statement 'mathematical object X is conscious' can be true in the hypothetical sense, as a logically true property of object X; but it does not commit to the existence of object X, and hence it is not necessarily true that there is anything which is conscious. Perhaps a physical object is needed which has structure X before it can be conscious.

This is an interesting variant of the ontological argument, which is fallacious for a similar reason. The statement 'A perfect God exists' is ambiguous between a hypothetical reading, that a perfect being would logically have to exist in order to be considered perfect, and an actual reading, that such a perfect being in fact does exist.

Is Tegmark's argument an instance of the 'ontological fallacy' described above? Unlike unicorns, a mathematical object such as the number two has an intricate set of properties that arise from its definition within a formal system. If a formal system is created which defines natural numbers and arithmetical relations, then an object such as '2' can have properties such as appearing in strings such as '2+2=4', being prime, having an irrational square root, and so on. But it is still difficult to avoid the conclusion that, without a computer to simulate this formal system, these strings remain only logical possibilities; the object '2' would have these properties if a computer were to simulate the formal system which defines it, but this is not the same as saying that there is a Platonically existing form '2' which has these properties. In other words, the hypothetical reading can hold and still leave room for reasonable doubt as to whether the actual reading holds.

To summarize, the main problem with Tegmark's argument is that someone can dissent from it by saying that, in the case of a computer simulating a universe, the crucial properties are that we have evidence that the information in the computer takes a definite

form which instantiates a conscious being; and one could go further and insist that it is the process of calculation which is special, as this is what makes a computer or a brain functionally different from a piece of paper or a memory stick, which merely can merely store the data to be read later.

How can one make a testable version of this question which can settle the matter? In one scenario, if one uses a computer to simulate a universe (or even just a brain which can interact with the outside world), then it will be conscious, while before the simulation, there will not be anything that is conscious. This is in principle an empirical question, because there either is or is not something there which is already experiencing the content of the simulation. It is potentially an ethical question as well, as simulating a brain which can suffer horrible experiences then becomes as unethical as subjecting a human being to the same experiences. Under the mathematical universe hypothesis, however, simulating the brain does not make any difference, as there is already a mathematical structure which is conscious and undergoing those experiences.

So although this question seems philosophical, it amounts to a big empirical and ethical difference in the outcomes. In one scenario, it is manipulation of information which creates a conscious entity. In the other scenario, it is the logical constraints on the way that information can be manipulated which create consciousness. In order to investigate the first scenario, various thought experiments have been tried by writers such as Daniel Dennett and Douglas Hofstadter to push the boundaries of what we are prepared to call conscious; for example, Hofstadter asks if a book containing a representation of Einstein's brain and instructions for manipulating information in that book would be conscious, and what would happen if someone were to make several copies of that book (Hofstadter and Dennett 1982). The fact that a human is performing a calculation means that something is created which experiences what Einstein would experience, including having his memories and thoughts. It does not matter if this task of calculating is distributed over several people or over a very long time, as the subjective experience of Einstein is independent of the computational time and who is performing the calculation. These scenarios may lead some people to conclude that the very premise is absurd: the act of calculating is not what creates a conscious entity, but merely creates a representation of it that we can see. Other people may be dogmatic in insisting that some calculation needs to take place. One version of Einstein is being simulated and other versions are not, and only the simulated version is conscious.

Tegmark uses the example of chess to illustrate this independence of the structure from what instantiates it. A particular game of chess has certain properties, such as being fifty-two moves long and using the Alekhine defence, no matter whether someone decides to play it, or how slowly, or over several people. Consciousness in this view is equivalent to properties of a chess game, which exist no matter whether someone is calculating it. This position, while easy to understand, could be disagreed with. Unlike the properties of a chess game, consciousness seems to be a property which we agree exists only if there is evidence of a computation taking place. The main advantage of the mathematical universe hypothesis is that is well-defined, while the alternative hypothesis is incoherent in not saying why consciousness should be any more than the mathematical properties which we observe in a computer simulation, or what independent existence or information mean. This hypothesis also suffers from a degree of anthropocentrism: computation is a property that humans have, but this does not mean that to explain it we need to invoke something behind the entire universe that is also capable of computation. Under the mathematical universe hypothesis, the fact that humans appear to be computing information is an artifact of the way we appear within our mathematical structure, which is in reality a static and immutable object rather than something changing (which is what a

computation would involve). The mathematical universe hypothesis suffers from a different problem, namely in being uncomfortably close to the ontological argument in conflating statements about possible properties of objects with statements that they actually exist.

The question of whether this new ontological argument is correct may end up being debated for a long time, as it is the primary obstacle to accepting Tegmark's argument. There may be an argument one day which resolves the conceptual confusion behind this debate, and points out clearly why these logical constraints are enough to create a conscious being – or conversely, why we need some notion of existence and information after all.

The main intellectual contribution of Tegmark's argument strikes me as something different from simply the philosophical question of whether a mathematical structure can be conscious. A good theory ought not only to answer fundamental questions, but should open up new domains of inquiry; such is the success of the theory of evolution, which not only answers fundamental questions about the nature of life, but opens up domains of inquiry into the family tree of living things and of different evolutionary processes. In a similar way, the understanding that universes are a type of mathematical object opens up new ways of investigating our own universe, and makes us aware of other universes that are equally worth exploring. We are capable of looking into alternative universes as rich in detail and as vast in scale as our own, which at least while being calculated will feel real to the entities being calculated, and hence have a claim to physical existence. To the extent that this idea can be realized, it will revolutionize every branch of science by giving us the means of exploring other universes, from their physics and chemistry, to their life forms and their evolution, and in some universes to the study of extra-terrestrial psychology and cultural history. The history of our own universe could even be explored, if we were able to create a formal system that could approximate it. Even if in practice impossible, we are in theory able to simulate the way that our planet has developed and the way that life has evolved, and even to investigate questions of human evolution and history from the origins of language to the causes of the French Revolution by testing alternative trajectories of these events, making them into rigorously controlled experiments.

The power of this idea, that formal systems can be universes in all their essential properties, makes almost irrelevant the question of whether they are different from 'real' worlds or not. It is possible that Tegmark is wrong and that mathematical universes differ from our own in not being simulated, but they would not be any less interesting to explore, and neither would their inhabitants be any less complex or capable of modeling and interacting with the world that they inhabit. While a complete theory of why we exist may therefore need to study the nature of the grand simulator behind our universe, this will be an addendum to what we are already able to explore using mathematics, namely the structure of laws that give rise to worlds containing self-aware entities. All of these possible worlds are mathematical objects waiting to be explored, requiring at most the power of reason to be called into physical existence.

References

D. Hofstadter and D. Dennett, *The Mind's I: Fantasies and Reflections on Self and Soul*. (Bantam Books, 1982).

M. Tegmark, The Mathematical Universe. *Found. Phys.* **38**,101-150 (2008).

M. Tegmark, *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. (Penguin Books, 2014).