The Diagonal Argument Strikes Again: Why Physical Systems Can Never Be Sure of Their Own Existence

Catherine M Reason[1]

London, United Kingdom

Chalmers has described the meta-problem of consciousness as the problem of understanding how and why we come to believe that we are conscious. Here we show that the meta-problem of consciousness is intimately related to another problem; the meta-problem of existence, or the problem of understanding how and why we come to believe that we exist. This problem is shown to lead to a version of Russell's paradox which makes it impossible for any physical system ever to be sure that it exists. The problem is illustrated by a thought experiment, the "sleepwalker paradox", which shows that no physical system can ever be sure that it is not in a dreamless sleep.

The Background to the Problem

David Chalmers has described the meta-problem of consciousness as the problem of understanding why we believe that we are conscious (Chalmers 2018). Exactly what is meant by "conscious" is not entirely clear. Some authors (see for example Churchland, 1996; Dennett 1996; Hacker 2010) appear to deny the existence of consciousness by any definition, while others (such as Frankish 2016, Blackmore 2016) apparently deny only the existence of phenomenal consciousness. As I understand it, the view advocated by Churchland, et al, which is usually referred to as eliminativism, is what Chalmers means by "strong illusionism", though I do not think the terminology is entirely clear.

There exists a little-known strand of work on the boundary between philosophy and mathematics, which has for some time been concerned with a related problem; that of determining whether any conscious physical system could ever actually be aware of being conscious. This approach was first outlined by Caplain (1995, 2000); criticized by Bojadziev (2000) and in a review by Dunlop (2000); and subsequently developed by Reason (2016, 2018). Since this approach assumes axiomatically the existence of consciousness in human beings, it does not bear directly on the meta-problem of consciousness. However it does apply to a closely related problem, which might be termed the *meta-problem of existence*, or the problem of understanding how and why we believe we exist. The traditional starting point for any such analysis is Descartes' epigram *Cogito ergo sum*, or "I think therefore I am" -- often known simply as the Cartesian Cogito. In what follows I shall show how Caplain's approach can be applied to the meta-problem of existence, and I shall illustrate what this tells us about the meta-problem of consciousness.

[1]Correspondence to: CMRneuro@Gmail.com

The Meta-problem of Existence

Let us assume that all our mental process supervene on one or more physical processes. The Cartesian Cogito can be represented as an inference as follows:

I think $\supset$ I am

The soundness of the inference obviously depends on the truth of the proposition "I think". Let us call this proposition p_0. This proposition must be established by a mental process we can express as the function r_0. Provided r_0 exists, then p_0 will be true, since r_0 is a mental process, and the proposition "r_0 is performed" can be regarded as equivalent to the proposition "I think".

How are we to determine, however, that r_0 is in fact performed? To do this we must establish the truth of the proposition "r_0 is performed", a proposition which we can call p_1. The proposition p_1 implies the existence of a mental function r_1 to establish the truth of p_1. So long as r_1 is performed, then p_0 is true, since the proposition "r_1 is performed" can also be regarded as equivalent to the proposition "I think". Once again, however, we can ask how it is possible to determine that r_1 is performed. In fact for any proposition p_n of the form "r_{n-1} is performed", there is an implied proposition p_{n+1} of the form "r_n is performed", which itself implies a mental function r_n to establish the truth of p_n.

It is apparent that in this way we can construct an infinite sequence of propositions together with the corresponding mental functions which establish them. How many of these functions actually need to be performed in order to establish that the proposition "I think" is certainly true? We can sidestep this question by defining a set R as being precisely that set of functions which must be performed in order to establish the truth of the proposition "I think". Since all our mental processes are assumed to supervene on one or more physical processes, this set of functions will have to be performed by some physical process, or set of processes, which we shall call X_0. A process will be regarded as a physical process if it has an *objective value*, which is to say that it has some property or properties which appear the same to all observers. We will define the *existence* of X_0 as such a property; X_0 exists if it can potentially be detected by all observers.

Provided that at least one of the functions in R is actually performed, the proposition "I think" will be true, since any proposition of the form "r is performed" is equivalent to the proposition "I think". So in order to establish the truth of "I think" we must establish that the proposition "At least one function in the set R is actually performed" is true. This proposition will be called (for reasons which will become clear later) $p_{diagonal}$. The function which establishes $p_{diagonal}$ we shall call $r_{diagonal}$. Since $r_{diagonal}$ must be performed if "I think" is to be established by X_0, $r_{diagonal}$ must be a member of R. If we are physical systems, then $r_{diagonal}$ will only be performed if X_0 exists. Likewise, any function or set of functions in R necessary for establishing that $r_{diagonal}$ is

performed must also be in the set R and will also only be performed if X_0 exists. We can therefore only be sure that $r_{diagonal}$ has been performed if we can be sure that X_0 exists.

How can we be sure that X_0 exists? There are potentially two ways of doing this, which might be termed the *rational* and the *empirical*. In the first case it might be possible to show that it is logically necessary for X_0 to exist. In the second case, even if the existence of X_0 is not logically necessary, it might be possible to demonstrate the existence of X_0 through observation. Such an observation would constitute a mental function, and clearly any and all such functions would have to be members of R. Since no member of R will be performed unless X_0 exists, we can only be sure that any such observation has actually been performed if we assume the existence of X_0. This makes the empirical method alone inadequate for establishing the existence of X_0, since it requires us assume what we set out to establish.

Therefore the existence of X_0 can only be established by logical means, if at all. But clearly, the existence of X_0 is *not* logically necessary -- if we ourselves did not exist then X_0 would not exist. And certainly there are possible worlds in which we do not exist. So there is no way to establish for sure the existence of X_0, and hence no way to establish for sure that at least one of the functions in R is performed.

We might postulate the existence of some other physical process, say X_1, which together with X_0, performs all the necessary functions. But this gets us nowhere at all, since we can simply redefine R as the set of all functions which must be performed by X_1 and X_0, in order to establish the truth of "I think". We can then define X to be the set of physical process containing X_1 and X_0, and the above reasoning applied to X_0 can then simply be applied to X. The same goes if we extend X to include any subsequent process X_n. Even if we allow the set X to be arbitrarily large, or of transfinite cardinality, the same reasoning applies.

We can summarize this reasoning as follows: if $r_{diagonal}$ is in the set R, then we can only be sure that $r_{diagonal}$ is performed if we first assume that X exists. But if $r_{diagonal}$ is not in the set R, then R cannot be the set of all functions which X must perform if X is to establish the proposition "I think". This is a classic Russellian paradox and the implication is quite simple: *the set of all functions which must be performed by any set of physical processes, if those processes are to establish the proposition "I think", does not exist.* In other words, there can be no set R for X to perform.

We can go further than this. Let postulate that the set of all functions which must be performed if we are to establish the truth of "I think" is a set R of cardinality N(R). We can then construct a set S, of the same cardinality as R, such that for every function $r_n \in R$ there is a corresponding element $s_n \in S$, such that for each r_n, the corresponding s_n is equal to1 if that function is performed, and equal to 0 if that function is not performed. There exists, in other words, a one-to-one correspondence between R and S, such that each element of S represents whether or not some given element of R has been performed.

Let us now construct the set S so that it consists entirely of zeroes. This corresponds to what in mathematics is called *diagonalizing* the set -- it shows that it is possible to

construct mathematically a state of affairs in which *no* function in R is actually performed. This gives us the diagonal proposition "No function in R is actually performed". This proposition must be shown to be false by some function, and if this function is in R, then the corresponding element in S can be set to zero. Therefore, this function cannot be in R. But R is by definition precisely that set of functions which must be performed to establish the truth of "I think"; so either the Cartesian Cogito is simply impossible, or there can exist no set S which can be put into one-to-one correspondence with R^2. This amounts to saying the cardinality of R does not exist -- that the set R is *larger* than any possible cardinality.

There are two immediate conclusions one can draw from this. Firstly, no physical system, and no conscious being whose mental functions all supervene on physical processes, can be certain that it exists, unless there is something fundamentally wrong with the human capacity for logical reasoning[3]. Secondly, if illusionism is to be used as a strategy for asserting the possibility of physicalism, then illusionism about consciousness is not enough; one would also need to be an illusionist about the Cartesian Cogito. And since the above reasoning applies equally well to alternative versions of the Cogito, in which the proposition "I think" is replaced by "I doubt" or "Thinking occurs", one would need to be an illusionist concerning *any and all* interpretations of the Cogito, and not just the usual Cartesian version.

So we can see that the meta-problem of consciousness is intricately connected to the more fundamental meta-problem of existence. It is our ability to determine with certainty that we, as living human beings, are performing functions of some sort -- whether those functions are perceptual, sensory, cognitive or introspective -- which gives rise to the meta-problem of consciousness.

The Sleepwalker Paradox

Some readers may find the idea of a system which is capable of reasoning, despite not being sure of its own existence, rather difficult to grasp. It might appear, on the face of things, that such a system would have to postulate that it did not exist in the possible world in which it exists. This however is a serious misreading of the situation, which arises from transposing a precise mathematical formalism into an ambiguous verbal representation. It would be a more accurate representation of the formalism, to say that such system would have to postulate that the possible world in which it existed was not the actual world.

We can illustrate this more clearly by considering, instead of the question "Can I be sure that I exist?", a simpler question: "Can I be sure that I am not in a dreamless sleep?". We shall define p to be the proposition "I am not in a dreamless sleep" and R

[2]More generally, one can say that there can be no surjective mapping from R on to any set of ones and zeroes, since any such set can be diagonalized. R is simply too large for any such mapping to be defined.

[3]It is obviously reasonable to point out that we cannot engage in any logical reasoning unless we exist. So one might object that we must exist, since we have engaged in logical reasoning. However this is simply the original problem in another form -- the proposition "I am engaged in logical reasoning" is clearly just another way of saying "I think". So evidently the same reasoning applies..

to be the smallest possible set of functions which must be performed to establish p. We shall define X to be any set of physical processes which can perform R.

We can now define the diagonal proposition $p_{diagonal}$ as follows: "All functions in R are adequately performed". It is clear that since R is precisely the minimal set of functions which must be performed in order to establish p, that we must be sure that $p_{diagonal}$ is true if we are to establish p; therefore the function which establishes $p_{diagonal}$ must be in R. But such a function will be adequately performed only if the set of functions R is adequately performed by X. Therefore if the function which establishes $p_{diagonal}$ is in R, we cannot establish $p_{diagonal}$ without first assuming $p_{diagonal}$. From which it follows that the function that performs $p_{diagonal}$ cannot be in R. So R cannot contain all the functions necessary to establish p. Since this conclusion contradicts the definition of R, we must conclude that there is in fact no such set as R.

This is the *Sleepwalker Paradox*, and it is clearly a variant of Russell's paradox. It is easy to see that, either we can never be sure that we are not in a dreamless sleep, or that we can determine that we are not in a dreamless sleep only by means of some non-objective process -- some process, that is, whose adequacy is not an objective property. Since we have already defined a physical process as a process whose properties are objective, this means that any process which tells us that we are not in a dreamless sleep cannot be a physical process.

There is clearly a close conceptual relationship between this notion of a sleepwalker, here defined as any being who is in a dreamless sleep but wrongly concludes that they are not, and the notion of a zombie, as defined by Chalmers (1996) and others. In fact every zombie is also a sleepwalker, although the reverse is not necessarily the case. It is also clear that there is a close resemblance between the Sleepwalker paradox and what Chalmers calls the meta-problem of consciousness; indeed, the Sleepwalker paradox can be regarded as just a formalization of the meta-problem.

What does all this tell us about the meta-problem? It tells us, firstly, that those whom Chalmers calls "strong illusionists" are required to commit themselves to the possibility that they (and everyone else) are in fact in a dreamless sleep. This is a precise requirement which cannot be evaded by equivocating over the meaning of the word "consciousness". Secondly, we can see the advantage of an approach to consciousness studies which relies on using precise, formal definitions to answer highly specific questions. We can also conclude that anyone who wishes to retain physicalism as a theory of human consciousness will also have to embrace strong illusionism, with all that entails. Doubtless other implications will become clear in future -- but for now, that is surely enough to be getting on with.

Discussion

Given that no purely physical system can be sure that it is not in a dreamless sleep, what sort of system, with what sort of properties, would be required to make such an assurance possible? We have already seen that the problem of determining the truth of "I think" can be represented as a set of functions which must be larger than any possible cardinality. A simple way of getting around the problem would be to allow

the set of processes X which perform R *also* to be of larger than any possible cardinality. In fact in principle this could even work for sets of physical processes, although there could never be enough matter in the universe -- even an infinite universe -- to constitute such a set of processes. However we should be careful to distinguish between mathematical representations and descriptive theories -- a mathematical representation involving an infinite set of functions does not necessarily entail an isomorphic theoretical description involving an infinite set of processes. Both the Sleepwalker paradox and the "meta-problem of existence" hinge on the point that any physical system must have *objective properties.* It is the objectivity of such properties as the accuracy and the existence of a physical process which makes the value of these properties a contingent, rather than a necessary fact. However one could axiomatically define a type of process whose accuracy and existence were necessary facts. Such a process would have to be intrinsically subjective -- it could have only those properties it was perceived as having by the observer who perceived it. This would require a way of understanding mental dynamics which is fundamentally different from the ways in which we understand physical systems. Any mental dynamics of this sort would also have to be intrinsically non-representational in nature. Ideas of this sort are likely to present a considerable conceptual challenge to researchers educated in a traditional cognitivist paradigm.

References

Blackmore, S. (2016). Delusions of Consciousness, *Journal of Consciousness Studies*, **23** (11-12), 52-64.

Bojadziev, D. (2000). Is consciousness not a computational property? *Informatica,* 24, 75-77.

Caplain, G. (1995). Is consciousness a computational property? *Informatica,* 19*,* 615-619.

Caplain, G. (2000). Is consciousness not a computational property? - Reply to Bojadziev. *Informatica,* 24*,* 79-81.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press.

Chalmers, D.J. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, 25 (9-10), 6-61.

Churchland, P. S. (1996). The hornswoggle problem. *Journal of Consciousness Studies,* 3 (5-6), 402-8.

Dennett, D. C. (1996). Facing backwards on the problem of consciousness. *Journal of Consciousness Studies,* 3 (1), 4-6.

Dunlop, C. E. M. (2000). Review of M. Gams, M. Paprzycki, and X. Wu (Eds) Mind Versus Computer: Were Dreyfus and Winograd Right? *Minds and Machines: Journal for Artificial Intelligence, Philosophy, and Cognitive Science* 10 (2), 289-296

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, 23 (11-12), 11-39.

Hacker, P. (2010). Hacker's Challenge. *The Philosopher's Magazine,* 51 (51), 13-32.

Reason, C. M. (2016). Consciousness is Not a Physically Provable Property. *Journal of Mind and Behavior,* 37 (1), 31-46.

Reason, C. M. (2018). A Theoretical Solution of the Mind-Body Problem: An Operationalized Proof that no Purely Physical System Can Exhibit all the Properties of Human Consciousness. (*In preparation)* -- preprint available at: http://arxiv.org/abs/1706.04192