The Journal of Mind and Behavior
Spring 2019, Volume 40, Number 2
Pages 95–120
ISSN 0271–0137

A No-Go Theorem for the Mind–Body Problem: An Informal Proof that No Purely Physical System Can Exhibit All the Properties of Human Consciousness

Catherine M. Reason

London, United Kingdom

This article presents an operationalized solution to the mind–body problem which relies on a well-defined effective procedure rather than philosophical argument. I identify a specific operation which is a necessary property of all healthy human conscious individuals — specifically the operation of self-certainty, or the capacity of healthy conscious humans to "know" with certainty that they are conscious. This operation is shown to be inconsistent with the properties possible in any meaningful definition of a physical system. I demonstrate this inconsistency by proving a "no-go" theorem for any physical system capable of human logical reasoning, if this reasoning is required to be both sound and consistent. The proof of this theorem is both general — it applies to any function whereby evidence affects the state of some physical system — and recursive, since any physical process subserving a function of this type is shown to imply another such function. Thus, for at least one aspect of human consciousness, the mind–body problem is resolved.

Keywords: interactive dualism, consciousness, Cartesian cogito

This is the second in a series of papers, which develop and formalize Caplain's 1995 argument, that consciousness cannot be a computational property. In a previous paper (Reason, 2016), I showed that Caplain's argument could be generalized from computational properties to all physical properties, and that this could be done without relying on philosophical concepts such as knowledge or belief. This paper however contained two significant weaknesses. The first weakness is that the argument relied on an assumption that all relevant mental processes can be represented as functions, and this assumption was never properly justified or made explicit. The argument also relied on an assumption that all healthy conscious human beings are capable of an operation described as self-certainty, or

Correspondence concerning this article should be addressed to Catherine M Reason, Institute of Mind and Behavior, PO Box 522, Village Station, New York, New York 10014. Email: CMRneuro@gmail.com

the ability to assert with certainty that they are in fact conscious. It is clear that many philosophers and cognitive scientists do not in fact accept this assumption.

In this paper, the reasoning behind Caplain's argument will be separated into two parts. Argument A from Reason (2016) will be expressed as the proof of a mathematical theorem, to the effect that no physical system capable of humanlike reasoning can assert that it is unconditionally certain of any proposition. This theorem will be called Theorem A, and its proof will be designated as Proof A, in accordance with the practice in Reason (2016).

I shall then prove a second brief lemma, to be called the *Cartesian lemma*, which will show how Theorem A can be extended to cover not only statements about consciousness but also statements about existence. This will show that, just as no physical system capable of humanlike reasoning can be certain that it is conscious, no such physical system can be certain of its own existence. This will place severe constraints on the viability of any materialist or physicalist theory of consciousness.

The logic behind Proof A will be straightforward, but the notation required to prove Theorem A to a satisfactory standard of rigor may seem dense and rather opaque at first sight. In order to emphasize the intrinsic simplicity of the proof, a verbal summary of the proof will therefore be included after the detailed exposition. Readers who are unused to following mathematical proofs may also find it useful to consult Reason (2016) before attempting to follow the detailed proof. Such readers should also bear in mind that there are significant differences between the proof of a mathematical theorem and the sort of arguments which are used to defend a philosophical thesis.

On the Difference between a Mathematical Proof and a Philosophical Argument

The difference between the types of arguments typically found in the philosophical literature, and a proof or derivation in formal logic, are summarized in an excellent article by Terence Parsons (1996). Since mathematical proofs have essentially the same structure as proofs in formal logic, Parsons' remarks may also be taken to apply to the differences between philosophical arguments and mathematical proofs. While proofs of this sort are rarely applied to problems in philosophy or cognitive science, readers from those disciplines should be aware that proofs are actually quite common in the natural sciences. In quantum mechanics, for example, it is common practice to show that some particular state of affairs is inconsistent with the basic principles of quantum mechanics by proving what is called a "no-go theorem" (see, for example, Bell, 1964 and Kochen and Specker, 1967). The proof in this paper can be regarded as an application of this method to theoretical psychology, rather than to theoretical physics.

To prove a no-go theorem, one must start by formalizing a set of axioms. By applying a given set of logical rules to these axioms, one can derive a conclusion.

Since this conclusion will necessarily follow from the set of axioms, any class of entities (say, theories or propositions) which satisfies that particular set of axioms will necessarily entail the same conclusion. A simple example from pure mathematics will serve to illustrate the point. In number theory the notion of the "size" of a number is formalized in terms of what is called the *cardinality* of sets. Any two sets are defined as having the same cardinality if the elements of one set can be put into one-to-one correspondence with the elements of the other. Any result proven for some particular cardinality will therefore also apply to any number which can be expressed in terms of that particular cardinality. For example, if one defines a set of cardinality *seven*, then any result necessarily true of *seven* will also be true of any other set of seven elements, since every set of seven elements can be put into one-to-one correspondence with the set *seven*.

It is important to note here a significant difference between the type of formal definition used in mathematical theory and the sorts of definitions used in, say, philosophy of mind or metaphysics. A philosophical definition is usually taken to be a description of the entity being described, and a satisfactory philosophical definition is therefore required to be as detailed as possible. A formal definition in mathematics, however, simply denotes a class of entities which satisfies that definition. The more general the definition, the larger the class of entities which it denotes. One can therefore see that from a philosophical perspective, a given definition may be simply vague; whereas from a mathematical perspective, the same definition may be sufficiently general to apply to an extremely broad class of systems, and that any result proved for such a broad class is extremely powerful. What is simply a weakness from a philosophical standpoint, therefore, can be seen to be a significant advantage from a mathematical one. This is the key advantage which the mathematical approach has over the philosophical one — the property of *generality*.

The mathematical method can be applied to the mind–body problem as follows. First, consciousness must be formalized in terms of an operation called self-certainty, which will be described more fully in a later section. This makes it possible to express consciousness in terms of a simple function (the function of answering a YES/NO question). Self-certainty is thus a specific example of a general class of functions, consisting of all functions which answer YES/NO questions.

If it can then be shown that, for a given class of physical systems, no YES/NO question can be answered with certainty, it will be clear that self-certainty is impossible in that class of physical systems. The necessary class of physical systems is that class of physical systems which can exhibit humanlike logical reasoning, since human beings must by definition exhibit humanlike reasoning. Any system capable of humanlike logical reasoning will be called *H*.

The proof which is to be presented here is the sequence of inferences in *H* which establish that no physical system reasoning in *H* can answer any YES/NO question with certainty. Therefore, if this proof is sound, no physical system capable of reasoning in *H* is capable of self-certainty. It follows that, if we assume that all healthy conscious human beings are capable of self-certainty, we have shown that no healthy human being can be an exclusively physical system.

The proof requires us to make two additional assumptions explicit:

1. Any mental process supervenes on some physical process;

2. No randomly selected physical process can be assumed a priori to perform any given function.

The proof can now be expressed as a couplet of statements:

By 1, the function of answering any YES/NO question must be performed by some physical process;

By 2, the accuracy of any physical process can be expressed as the function of answering a YES/NO question such as "Is this process randomly selected?"

This couplet clearly generates an infinite regress. This is the essential outline of Proof A. Because the proof can be summarized as a couplet of sentences, I have elsewhere sometimes referred to Theorem A as the *Gemini theorem*. Much of the rigorous form of Proof A will be devoted to showing that every YES/NO question satisfies the mathematical definition of a function.

The Concept of Self-Certainty

It is unnecessary to become embroiled in metaphysical speculations about the nature of consciouness. Consciousness itself will therefore be formalized in terms of the operation of *self-certainty*, defined as follows:

Definition: *Self-certainty* is the capacity of at least some conscious beings to verify with absolute certainty that they are conscious — that is, to give the answer YES to the question "Am I certainly conscious?"

It is important to emphasize that it is only *absolute* certainty which is at issue here. Caplain's proof does not imply that physical systems cannot be *reasonably* certain of being conscious, or that they might have a certainty of being conscious which is contingent on additional assumptions. For example, Chalmers (2012) postulates an analogous situation involving mathematical reasoning; in this example an expert mathematician is given cause to doubt the accuracy of his own mathematical musings by being offered the possibility that he has been administered some sort of drug which destroys the capacity for mathematical reasoning. Chalmers suggests that the mathematician could regain confidence in the accuracy of his own cogitations by insulating himself from all such questions (that is to say, by ignoring them). Certainty of this sort, which depends on additional assumptions, is not covered by Caplain's proof and does not constitute self-certainty. We can express this as follows:

Condition 1: An entity is self-certain only if its certainty is absolute (beyond any possibility of error) and this certainty does not depend on additional assumptions.

It is also necessary to emphasize that self-certainty does not require assumptions to be made about the nature of consciousness. Any such assumptions — for example, that consciousness equates to wakefulness, or that consciousness requires some specific sort of self-awareness — should therefore not be made. A conscious being might, for example, be experiencing a lucid dream or some other altered state of consciousness. So long as the conscious being can verify that it is in some conscious state or other, the requirement for self-certainty is satisfied. We express this as:

Condition 2: Self-certainty does not require that the conscious state which is found to exist has any particular property or set of properties.

The significance of this condition cannot be overstated. It has repeatedly been pointed out to me that nothing can be inferred about the nature of consciousness from such formalized definitions. Condition 2 illustrates that this situation is intentional. The no-go theorem to be proved in this paper depends on this formal definition alone, and not on any philosophical notions about what consciousness is or should be. To underline this, one can express self-certainty as *illusory self-certainty*, as in the sentence "It is absolutely certain that I have at least the illusion of being conscious" without affecting the validity of the proof.

Assumptions Necessary for the Proof of Theorem A

In order to express the proof in a form that can reasonably be described as rigorous, it is necessary to list the assumptions and premises on which the proof depends, in a clear and unequivocal form.[1] Let us start by defining a physical system as some set of physical processes, where a physical process is defined as follows:

Definition: A *physical process* is any entity which has an objective existence and is capable of evolving in time.

[1] The premises listed here were originally given in Reason (2016). They are presented here in an expanded form, together with explanations as to why they have been chosen.

I have made no attempt here to explicate what is meant by the terms *objective existence* and *time*. If necessary, one can define objectivity in terms of some set of observers, whose observations of some given physical process are related by some transformation group — say, the Lorentz group. Time can be regarded as just some parameter which governs the evolution of such a system. In the future it may turn out to be necessary to explicate these terms further; at the moment, however, I believe such a degree of elaboration is simply unnecessary and beyond the scope of this paper.

Philosophers use the notion of supervenience to refer to entities which are not themselves physical systems but whose existence is nonetheless consistent with the doctrine of physicalism. An entity S is said to be supervenient on some basis B if two things which have the same B properties must have the same S properties (Kim, 1984). Mental processes are thus assumed to supervene on physical processes. Hence the following expression:

Principle of Physicalism: All mental processes supervene on some physical process or set of processes.[2] (The notion of a *mental process* will be expressed more formally in the next section.)

The next two premises refer to the rules of logical inference by which Caplain's proof can be established. In the natural sciences, no-go theorems are usually proved within the context of some axiomatized formal system which corresponds to a theoretical description of the natural system under study. However, no such axiomatized theoretical description exists in psychology or the philosophy of mind. It is therefore necessary to adopt a more indirect approach.

To perform any sort of mathematical or logical reasoning, it is necessary to make the assumption that we as human beings are capable of doing so. Any system which incorporates the human brain's capacity for logical inference, we shall call *H*. This leads to the following two assumptions:

Assumption 1: There exists a physical system M which supports the properties of *H*;

Assumption 2: *H* is in principle both sound and consistent.

"Sound in principle" simply means that if a system reasoning with *H* makes a mistake, there is no reason in principle why it should eventually not correct that mistake.[3] *H* does not therefore condemn any system using it to fallacious logical

[2] Referred to as *Principle F′* in Reason (2016).

[3] Another way of looking at this, which is perhaps strictly more accurate, is to assert that *H* is sound but that individual intelligent beings are noisy theorem-provers in *H*. That is to say, intelligent beings can prove theorems in *H* but are subject to random errors. However, such beings are not subject to *systematic* errors.

reasoning which can never be corrected. In practice, the condition that *H* must support humanlike reasoning requires *H* to support at least a capacity for arithmetic and classical modal inference. There may well be many different formal systems, with different sets of axioms and different rules of inference, which exhibit the necessary properties of *H*, or it may be that there is no such formal system. It is extremely important to emphasize at this point that *H* must not be *assumed* to be a formal system — *H* is simply the label we give to any system which supports a capacity for humanlike reasoning, regardless of whether or not it can be fully axiomatized.[4] However since *H* is defined only implicitly, proofs in *H* cannot be given in the way they would be for formally specified systems. Such proofs must be presented instead in terms of the classical modal logic and arithmetic which supervene on *H*. For this reason the following proof must be presented in considerable detail, so that the validity of the proof is clear to the reader. The proof to be found in the following section can be regarded as an *effective procedure* for proving the Theorem A, which can be followed mechanically without any special ingenuity or insight; we shall examine in more detail exactly what is meant by an effective procedure in the final section of this paper.

A Rigorous Formulation of Caplain's Proof

We are now in a position to present Caplain's proof in a rigorous form. The proof will now be presented in a number of stages, and each milestone will be clearly identified and numbered. The reader is invited to consider the reasonableness of each inference as it is presented. Should any reader find the accompanying profusion of Greek letters and other algebraic symbols somewhat bewildering, they may find it useful to examine the verbal summary of the proof in the following section, in which these symbols do not appear. However, readers should be aware that the notation presented here is necessary for a rigorous demonstration of the proof.

It will be convenient to restrict our considerations to mental processes which can be represented as functions.[5] We will define M to be some physical system, and φ to be some function which yields the correct answer to a YES/NO question. We shall begin with the simplest case in which all questions can be considered to refer to propositions which are either TRUE or UNTRUE. The function φ therefore represents the mapping:

[4] Since *H* cannot be assumed to be a formal system, it is necessary to understand the concept of soundness in a similarly informal way. A deductive system (whether or not it is fully axiomatizable) can be regarded as sound just so long as it is impossible in that system to prove a false conclusion from true premises.

[5] I use the term "function" here in its mathematical, or declarative sense, not in the imperative sense which is sometimes used in computer science. That is to say, a function is understood to be an abstract mapping between two sets, and not a specified operation to be applied to some datum input.

$$\text{TRUE} \rightarrow \text{YES}$$
$$\text{UNTRUE} \rightarrow \text{NO}$$

where TRUE and UNTRUE are the possible truth-values of some proposition p. We shall call the range of this function the *evidence state* of p. So if T represents the truth-value of p, and K represents the evidence state of p, the mapping can be represented as:

$$K = \varphi(T)$$

We now have our first milestone:

Milestone 1: Any mental process which is equivalent to the process of correctly answering a YES/NO question satisfies the formal definition of a function of the form φ.

The equivalence covers mental processes which may not themselves be answers to YES/NO questions, but can be represented as such. For example, the mental process of thinking "X is true" is equivalent to the process of answering the YES/NO question "Is X true?" In general, any proposition whose truth value maps onto some given evidence state of M, can be represented as the answer to some given YES/NO question.

In a physical system M, an evidence state of some proposition p can be any state of M correlated with the correct truth-value for p, and the function φ will be performed by some physical process P, as is required by the Principle of Physicalism. This gives us a general format for mental processes in a physical system. If T is the truth value for some proposition p, and K is the corresponding evidence state for T, then:

$$K = O(P)$$

where O is the output of some physical process P. This is to say that if TRUE is the correct truth value for p, then M will evolve to state K_T, and if TRUE is not the correct truth value for p, then M will evolve to some state nonK_T. Thus P executes the mapping:

$$\text{TRUE} \rightarrow K_T$$
$$\text{not TRUE} \rightarrow \text{non}K_T$$

where K_T and nonK_T are separate states of M. We shall call this mapping π.

It is possible that T is a physical state of M, and furthermore that T is the same physical state as K; K_T however cannot be the same physical state as nonK_T. Some decision is therefore always required as to whether M evolves to K_T or to nonK_T. This decision, and the subsequent evolution of M, constitute the process P. This leads us to our second milestone:

Milestone 2: Any physical instantiation of a function in the form of φ can be expressed as a mapping in the form of π, which will be performed by some physical process P.

Now consider the mapping which does the *reverse* of π; that is, the mapping:

$$\begin{array}{l} \text{TRUE} \rightarrow \text{non}K_T \\ \text{not TRUE} \rightarrow K_T \end{array}$$

a mapping which we shall call ρ. It is nonetheless clear that ρ has the same *form* as π, and by Milestone 2 there will be some physical process which performs it, which we shall call R. We shall express this as our next milestone:

Milestone 3: For every mapping of the form π performed by a physical process P, there will be some contradictory mapping of the form ρ and some physical process R which performs this.[6]

But if K represents the evidence state of some true proposition, how is M to ascertain whether the state K has been produced by P or by R? It clearly matters, because if K has been produced by P then K_T will represent the answer YES to some question, but if K has been produced by R, then K_T will represent the answer NO to that same question. To resolve this difficulty, it is clearly necessary for M to ascertain whether K has been produced by P or by R!

Another way of looking at this is to say that P is an *accurate* instantiation of φ but that R is an *inaccurate* instantiation of φ. In this case K_T will always represent the evidence state YES, but if K has been produced by P then K will be a correct evidence state, whereas if K has been produced by R then K will be an incorrect evidence state. This brings us to the notion of *failure*, which is always a necessary consideration when dealing with processes in the real world, as opposed to the idealized mathematical functions they perform. We can express this notion in terms of the following axiom:[7]

[6] This can always be done by applying a logical NOT operator to the truth value before performing φ. Therefore if P exists, so does R.

[7] It is convenient at this point to express the notion of failure as an axiom of *H*, but it is not strictly necessary. Later I shall show how it is possible to express the notion of failure without assuming the Axiom of Fallibility. In the meantime, the Axiom can be understood intuitively by imagining a bag full of pebbles, each of a different color: the Axiom asserts that if one were to take a pebble from the bag, one could have no way of ascertaining the color of the pebble without looking at it, or examining it in some way.

Axiom of Fallibility: Given any arbitrarily selected physical process, whose properties have not been ascertained, it is impossible to be certain that this process will or will not perform some given function.

In other words, if we have a physical system M in which there exists some evidence state K, produced by some process which is supposed to instantiate a function of the form φ, it is impossible to be sure whether that state has been produced by the process P or by the process R without ascertaining any of the properties of the process. We might be able to deduce through logical deduction what that process should be, but we cannot through logical deduction alone deduce what that process actually *is* in any given case.

How could we ascertain the properties of this mysterious process? We could either examine the process directly, or we could examine the mapping performed by the process, which entails examining the states between which the mapping occurs. In the former case we could ask, for example, "Does this process have the properties of P or of R?" This is clearly equivalent to asking either the questions "Does this process have the properties of P?" and "Does this process have the properties of R?" These are clearly YES/NO questions and so by Milestone 1 can be expressed as functions of the form φ — let us call this particular function $φ_1$. In the second case we could ask, for example, "Does this process map T on to K?" This is also clearly a YES/NO question and so is also expressible as a function of the form φ. Let us call this particular function $φ_2$.

By Milestone 2, any physical instantiation of these functions can be expressed as a mapping of the form π, which will be performed by some physical process. In either case call this process P'. By Milestone 3, there will exist for this mapping a corresponding mapping of the form ρ, which will be performed by some physical process we shall call R'.

From the Axiom of Fallibility it follows that one cannot determine if the *actual* physical process which instantiates $φ_1$ or $φ_2$ in the physical system M, is P' or R' (or indeed some completely different third process as yet unidentified) without ascertaining the properties of this actual process. But ascertaining the properties of an actual physical process (as opposed to an idealized one) is equivalent to a function of the form φ. And by the Principle of Physicalism, M can only perform functions of the form φ by means of some physical process. This leads to the following rather awkward state of affairs:

1. A physical system M can only perform some function of the form φ by means of some physical process;

2. M can only ascertain the correctness of some physical process by performing some function of the form φ.

This is either circular or leads to an infinite regress. The nature of the infinite regress is easily demonstrated by examining the functions which need to be performed. If the first function can be represented as:

$$K_0 = \varphi_0(T_0)$$

where K_0 is the evidence state of some proposition p_0 and T_0 is the truth value of p_0; then subsequent functions can be represented as:

$$K_i = \varphi_i(T_i)$$

where K_i and T_i are the evidence state and truth value respectively of some proposition p_i where p_i is equivalent to answering the YES/NO question:

"Is the function which answers the question equivalent to p_{i-1} correctly performed?"

Clearly the number of questions which would need to be answered in order to establish that any given φ-type function has been performed correctly is at least countably infinite.[8] Let us now assume that there exists some countable infinite list L_1 of processes, which can be put into a one-to-one correspondence with a list L_2 of φ-type functions, such that each φ_i is mapped onto the physical process which performs it. We can then construct a function φ_{omega} equivalent to the question "Is it not the case that every process in L_1 is in fact a process of type R?" Clearly φ_{omega} cannot be on the list L_2 since it is not a function of the form φ_i. Therefore the set of questions which would need to be answered in order to establish that any given φ-type function has been performed correctly is *more than* countably infinite, as is the set of physical processes needed to perform those functions. In fact for *every* set of functions, however large, which is in one-to-one correspondence with the physical processes which perform those functions, there will be some function $\varphi_{diagonal}$ equivalent to the question "Is it not the case that every process in that set has been performed by a process of type R?" which is not in that set.[9]

The upshot of all this is that M can never establish with provable accuracy whether any of its processes is accurate or inaccurate, since establishing this requires a non-terminating sequence of processes.[10] It does not yet follow, however, that M cannot arrive at the conclusion that all its processes are accurate. M might, indeed, be correct to conclude that all of its processes are accurate.

[8] In mathematics, the term "countably infinite" refers to infinite sets whose elements can be put into one-to-one correspondence with the natural numbers.

[9] This is an example of what in mathematics is called the *diagonal method*.

[10] It would strictly be more accurate to say that the accuracy of M cannot be proven in the logical system *H*.

However, such a conclusion cannot be a logically valid theorem in *H*; there is no logically sound chain of reasoning in *H* by which M can show that any one of its processes is accurate. To put this in slightly less technical language, it is impossible to prove in *H* that any of M's processes is accurate.

If this is not yet obvious, consider the following: we have established above a chain of reasoning which shows that any attempt by M to establish the accuracy of any of its processes logically implies an infinite regress. We have established this conclusion by means of our human capacity for logical reasoning; a capacity which we have labeled *H*. By Assumption 2, *H* is in principle sound and consistent. If *H* is sound and consistent, then it cannot support a stable chain of reasoning which would lead to a conclusion inconsistent with other conclusions already established in *H*; that is to say, with other theorems in *H*.[11] M might be able to derive a logically valid proof in some other system, say *nonH*, that some of its processes were accurate; but then by definition M could not be a human being.[12] One can therefore say that the statement "M cannot prove the accuracy of any of its processes" is equivalent to a theorem in *H*.

If at this point we invoke Assumption 1, then M itself can establish that none of its processes is provably accurate. Since M can reason according to the system *H*, M can prove any theorem that can be proven by a human being; this includes the proposition "M cannot prove the accuracy of any of its processes." To express this in slightly less formal terms, we can say that if a human being, reasoning according to the principles of *H*, can show that M cannot establish the accuracy of any of its own processes, then M can do the same.

This is an extremely raw and primitive version of the proof we need, which applies only to propositions which are either categorically true or categorically untrue. In real life, we are often concerned with degrees of probability as much as simple truth or untruth. Often we will have to deal with questions where we simply do not know the answer; where the possible answers to some question will be YES, NO, and MAYBE. We can extend the formalism to include such questions by representing them in the form:

[11] Since *H* is only sound and consistent in principle, it could temporarily support an invalid chain of reasoning inconsistent with other theorems in *H*. However, such a chain of reasoning would have to be unstable, by which I mean that *H* would also have to support a chain of reasoning showing this fallacious chain to be invalid.

[12] Such a system might, for example, contain axioms which simply assert that certain physical processes are accurate. It would be necessary for such a system to be so constituted that these additional axioms did not lead to contradictions with other properties of the system; this would entail *nonH* abandoning certain axioms, such as the Axiom of Fallibility, and having different rules of inference from *H*. The rules of *H*, for example, allow M to ask the question "Is there any reason why these additional axioms should not be dropped?"

"Can M be certain that some proposition p is certainly true?"

All questions of this type can be expressed as YES/NO questions, in which the MAYBE category simply collapses into the NO category. We can now encapsulate the reasoning in this section as a specific proof, which we shall call Proof A. This proof establishes that no physical system which possesses reasoning capacities equivalent to those of a human being can correctly answer YES to the following question:

"Can M be certain that it is answering some given YES/NO question correctly?"

We can express this as our next milestone:

Milestone 4: For any suitably qualified physical system M (that is, one which is capable of human-level reasoning), and for any proposition equivalent to the answer to a YES/NO question, there can exist some physical process of type R which will answer that question inaccurately. Therefore any suitably qualified physical system can deduce that it can never be certain that its answer to any YES/NO question is correct.

This is an extremely general result and it is thus worthwhile checking for any possible flaws. The reader might, for example, consider that since Proof A applies only to physical systems, it might not apply to M if M is not certain that it is a physical system. We can deal with this by expressing Milestone 4 in a conditional fashion:

"If M is a physical system there can always exist some physical process of type R which will answer any given YES/NO question inaccurately."

Clearly this statement is provable in *H*. In order to establish that Proof A certainly does not apply to M, M must be able to answer YES to the question:

"Can M be certain that M is not a physical system?"

We will call the proposition equivalent to this question p_0. M can now ask the question p_1:

"Is there some process R_0 in M that would answer p_0 incorrectly?"

Whatever the answer to p_1, M can now ask the question p_2:

"Is there some process R_1 in M that would answer p_1 incorrectly?"

Clearly we find ourselves in another infinite regress. In fact, generally for any proposition p_i which can be expressed as a YES/NO question, there will be some proposition p_{i+1} of the form:

"Is there some process R_i in M that would answer p_i incorrectly?"

Thus the sequence of questions which need to be answered in order to establish the truth of the proposition:

"M can be certain that it is not a physical system"

is non-terminating. Since M is in fact a physical system, each such question is equivalent to a φ-type function which must be performed by some physical process, which means the sequence of such processes is also non-terminating. So long as M is in fact a physical system, it does not matter if M itself is not certain that it is a physical system.

Having established this, our first application of Proof A is to the question:

"Can M be certain that M is conscious?"

where this question is subject to Conditions 1 and 2 as described in the previous section. Proof A shows that no suitably qualified physical system can answer YES to this question. Since *H* is required to be consistent it therefore follows that no suitably qualified physical system can answer YES to the question "Can I be certain that I am conscious?" Any healthy, conscious human being should be able to answer YES to this question. Conscious human beings, therefore, cannot be exclusively physical systems.

It is important to note here that Proof A applies to *all* φ-type functions. It therefore applies to all functions which answer questions of the form: "Can I be certain that I am conscious?" since all such questions form a subset of φ-type functions. The importance of Conditions 1 and 2 now become apparent: it is irrelevant what "consciousness" is understood to mean, since regardless of what it is taken to mean, all questions referring to it must be answered by φ-type functions which are subject to Proof A.

Before proceeding to the next section, we will now deal with the outstanding matter of the Axiom of Fallibility. The purpose of this axiom is to express the concept of failure, by asserting that there is no way of establishing through purely logical means whether some particular physical process is working correctly or incorrectly. Some readers, however, and most especially philosophers, may be attracted to the idea of simply dropping this axiom. One can, however, deal with the notion of failure without expressing it explicitly as an axiom. Consider some function of the form φ which is performed by some physical process. M can now ask the question:

"Is it certain that the process which was supposed to perform the function φ did not in fact perform a mapping of the form ρ?"

to which the answer must be YES. The process of answering this question is itself a function of the form φ — call this φ'. By the Principle of Physicalism, any physical system which has performed this function must have done so via some physical process. M can now ask the question:

"Is it certain that the process which was supposed to perform the function φ' did not in fact perform a mapping of the form ρ?"

to which the answer must be YES. Since the process of answering this question is also a function of the form φ, which we can call φ'', it must by the Principle of Physicalism have been performed by some physical process. In fact, in general any function φ_i which is assumed to have been performed correctly implies the existence of another function φ_{i+1} which answers the question equivalent to that assumption. This implies a non-terminating sequence of functions and a corresponding non-terminating sequence of physical processes to perform them. There is, in other words, no escape from the infinite regress by abandoning the Axiom of Fallibility.

The Cartesian Lemma

A special problem exists in the case of self-referential statements such as the question "Is it certain that I exist?" In such a case a system reasoning in *H* could get as far as Milestone 2 and then argue, along the lines of Descartes' *Cogito ergo sum*, that it does not really matter if the physical process which subserves the answering of that question is accurate or not — the mere fact that such a physical process exists is enough to answer the question in the affirmative. If M is asking "Does M exist?" then M exists. Descartes' *Cogito* can be represented in *H* as a syllogism of the form:

If I think then I am;
I think;
Therefore I am.

Substituting for terms, this becomes:

If M is asking "Does M exist?" then M exists;
M is asking "Does M exist?";
Therefore M exists.

However this syllogism requires M to establish that it is the case that M is asking "Does M exist?" By Milestone 1 of Proof A, this question entails a φ-type function. This inference enables us to state the following principle:

Cartesian principle: In order to establish that M is asking some question, M must perform a φ-type function.

Since the process of establishing if M is asking "Does M exist?" implies a φ-type function, Proof A (including the Cartesian lemma) will apply to it, as before. This can be expressed as an intermediate milestone in Proof A.

Milestone 1a: No physical system capable of reasoning in *H* can be certain that it is performing any φ-type function without performing some other φ-type function.

This obviously generates a non-terminating sequence of φ-type functions, which implies a second intermediate milestone:

Milestone 1b: No physical system capable of reasoning in *H* can ever be certain that it is performing any φ-type function without performing an infinite sequence of φ-type functions.

By Milestone 2 from Proof A, any such sequence must be performed by some physical process or set of processes. Let us use the symbol S to refer the set of functions performed by some physical process P. We can now ask the question:

"Does P exist?"

Plainly if P does not exist then no function in S will actually be performed. We can construct a set S' which is in one-to-one correspondence with the set S, such that for every φ-type function in S which is actually performed, there is a corresponding element 1 in S', and for every φ-type function in S which is not performed, there is a corresponding element 0 in S'. Using simple diagonal reasoning we can now construct a set S' in which every element is 0, from which it follows that every function in S maps on to an element 0 in S'. By Milestone 1 of Proof A, the question "Does P exist?" can be represented as a φ-type function. We now consider whether this function is in the sequence S, given that it is logically possible for every element in S to map on to an element 0 of S' — in other words, it is mathematically possible that no function in S is actually performed. If the function implied by the question "Does P exist?" is in S, then it is logically possible that no such function has been performed. However if there is no such function in S, then no such function can have been performed by the process P, since S is precisely that set of functions performed by P. Since this is true generally for any set S, no matter how large, we can state the following lemma:

Cartesian lemma: No physical system capable of reasoning in *H* can ever be certain that it has performed any φ-type function.

Some philosophers assert that the *Cogito* is not, in fact, a syllogism at all but a sort of intuition which should be expressed in the following form:

M is asking "Does M exist?" and therefore M exists.

However such an intuition still requires M to establish that it is the case that M is asking "Does M exist?" Once again, this entails a φ-type function to which Proof A will apply.

Verbal Summary of the Proof

Proof A has two important properties. Firstly, it is generalizable; it applies to any mapping of the form φ. Secondly, it is recursive; it shows that the question of whether any given φ-type function is performed correctly can itself be represented as a φ-type function, to which Proof A applies.[13]

We shall now summarize the proof given in the last section in a somewhat more accessible, verbal form. In doing so we shall illustrate the properties of generality and recursiveness, and show how these can be used to deal with the apparent objections to the proof which have been raised informally in correspondence to the author. It will be useful to base our verbal summary on the notion of *evidence*; however, we shall ignore all considerations about the nature of evidence and instead operationalize the *function* of evidence as a sequence of YES/NO questions. That is, if M has evidence that some proposition is true, then it is the function of that evidence to answer the question "Is this proposition certainly true?" with either a YES or a NO. This function we shall refer to as the *evidence function*, which relates the presence or absence of some evidence to the evidence state K of some proposition. If we define a state E so that E is 1 when some evidence exists, and E is 0 when that evidence does not exist, then the evidence function relates E to K as follows:

If $E = 1$ then $K = \text{YES}$
If $E = 0$ then $K = \text{NO}$

[13] These properties of generality and recursiveness are explicit in the structure of Proof A, whereas in Argument A from the author's previous paper (Reason, 2016) they are not. This can be seen most clearly by comparing the structure of Argument A with the Gemini couplet at the end of this section.

We shall assume that the evidence for some proposition is always perfectly correlated with the truth value of that proposition (since it is trivially obvious that evidence which is not so correlated cannot support certainty).

By the Principle of Physicalism, any evidence function in M must be performed by a process which supervenes on some physical process or other. Since we have so far ascertained nothing about this process, we shall call it X.

Philosophers, and most particularly epistemologists, traditionally use possible worlds semantics to describe conditions of possibility and necessity (see, for example, Kripke, 1963). For convenience we shall adhere to this convention henceforth; if some proposition is possibly true, we shall express this by saying there exists some possible world in which that proposition is true. We shall refer to a possible world as *epistemically available* to M if M cannot rule out the possibility that M exists in such a world, given the evidence currently available to M. We shall say that M is certain of some proposition if M can correctly decide that there is no possible world *epistemically available* to M, in which the evidence state which corresponds to the truth of that proposition is itself incorrect.

We shall express the first part of our proof as the following logical argument, which we shall call Castor:

It follows from the Principle of Physicalism that any evidence function must supervene on some objectively real process X. Since X is an objectively real process it cannot be guaranteed a priori to be a correct instantiation of any function. We shall say that X *fails* on any occasion on which it does not correctly map some element from the domain of the function to its range. If X fails than its output will be an incorrect evidence state.

How can M be certain that the evidence state K is correct — that is, that the mapping from E to K has been correctly performed? Only by ruling out the possibility that X has failed. In other words, there must be no possible world epistemically available to M in which X has failed. The question of whether X has failed can be represented as the question “Is it certainly the case that X has not failed?” The process of answering this question is itself equivalent to a new evidence function.

This argument is general — it applies to any evidence function. We can express this as the second part of our proof, which we shall call Pollux:

Castor has generality — that is, it applies to any evidence function. Therefore Castor can be applied recursively to the new evidence function generated by Castor. This will generate a third evidence function, to which Castor can be applied recursively yet again. Indeed each time Castor is applied to any evidence function, it will generate another evidence function to which Castor can be applied. Therefore the correctness of any objectively real process X cannot be ascertained without performing a non-terminating sequence of functions, each of which (by the Principle of Physicalism) must be performed by some objectively real process.

Castor and Pollux together are equivalent to the Proof A given in the previous section. By splitting the proof into two sections in this way, the reader can clearly see how Proof A exhibits the two useful properties of generality and recursiveness. Castor establishes generality; Pollux shows that Castor can be applied recursively.

The foregoing tells us that M can never guarantee its own correctness when evaluating the truth of any proposition. However, this result has not been proven by M, but by us, the author and readers of this paper. It has been proven in whatever logical system or systems underly our reasoning processes, which we must assume are sound in principle. As in the previous section, let us call this system *H*. However Assumption 1 allows M to "inherit" *H*, as it were, and so to prove Proof A for itself. This reasoning also allows us to disregard the detailed character of *H* itself, since these details effectively cancel out. The reader's understanding of Proof A can effectively be regarded as a derivation of Proof A in the logical system *H*.

We shall now express the outcome of Proof A as a theorem, which we will call *Theorem A*:

No suitably qualified physical system can exhibit the property of self-certainty, and in any system which exhibits self-certainty, the process which subserves self-certainty cannot depend on any objectively real process.

Readers should note that Proof A is a proof of exactly this theorem, and not any other. Some readers of my previous paper have been tempted to paraphrase this theorem into a generic and rather ambiguous statement about the epistemology of consciousness, and then to express objections based on the resulting ambiguities. However, a theoretical proof is a proof of a specific, well-defined proposition, derived from a specific set of assumptions according to clear rules of inference. In this respect a theoretical proof is quite different from a philosophical argument.

Proof A can be expressed as a couplet of statements, as follows. For any suitably qualified physical system:

1. By the Principle of Physicalism, any evidence function must be performed by some physical process;

2. By the Axiom of Fallibility, the correctness of any physical process must be determined by some evidence function.

All philosophical objections to Proof A can be dealt with by applying this *Gemini couplet*. We shall examine some of these objections in the following section.

Discussion

The reader may well now be wondering, if Proof A is as generalizable as it appears to be, why it does not simply apply to *all* processes, whether physical or not. What sort of characteristic could possibly differentiate physical processes, which are subject to Proof A, from non-physical processes, which are not subject to Proof A?

The answer is that physical processes are defined as being objectively real — that is, they are assumed to have properties which exist independently of the subjective states which record the values of those properties. Therefore if we assume a φ-type function performed by some objectively real process X, there will be an objectively real fact about whether X is a process of type P, or a process of type R. The infinite regress, in the form of the non-terminating sequence of processes generated by Proof A, arises from the need to identify exactly what this objective property is. No such difficulty arises if no objectively real process exists, since there is then no objective property to be identified. The consequence of this, however, is that whatever process performs the φ-type function equivalent to self-certainty, this process cannot depend on any objectively real process — it cannot, in other words, depend on any process which is external to human consciousness. Another way of looking at this is to say that any process which performs self-certainty must be subjectively real but not objectively real.

Theorem A itself shows only that self-certainty is impossible in any physical system. In order to apply this to the mind–body problem we need to make a further assumption, which is that self-certainty is in fact possible for human beings. This is by its nature an empirical question. Some will regard it as intuitively obvious that human beings are capable of self-certainty, given that self-certainty does not require us to make any assumptions about the nature of consciousness or the properties of conscious states. This is especially the case if Proof A includes the Cartesian lemma, since the Cartesian lemma shows that if human beings were physical systems, we could never be certain that we were not in fact dead (in the sense of completely non-existent, as opposed to preserved in some sort of virtual reality afterlife). To the best of my knowledge, no philosopher or cognitive scientist has been prepared to endorse such a proposition.

For any readers who are prepared to consider such an extreme case, however, there is a clear empirical prediction which can be made. Theorem A implies that self-certainty will entail a violation of the conservation of energy (as was explained in Reason, 2016). We can express this violation as an inequality between the energy liberated within a physical system and the energy dissipated by that system. The difference between energy liberated and energy dissipated is denoted by the Greek letter χ (Reason, 2016). This inequality could in principle be detected experimentally; such a topic is however beyond the scope of this paper.

Let us now examine some common objections which have been raised against Theorem A and its implications. One objection is that conscious entities do not need to calculate or prove that they are conscious, but can in some way simply "ascertain" it. The mistake here is to assume that physical systems must necessarily have the same properties as conscious human beings. Any such "ascertaining" can still be represented as an evidence function, and if such an evidence function is performed by some physical process X (as is required by the Principle of Physicalism), then Proof A will apply to it. Readers should be careful not to attribute unconditionally to physical systems the properties of their own consciousness. Proof A applies generally to any system capable of reasoning in *H*, if both the Principle of Physicalism and the Axiom of Fallibility are assumed.

A further objection which has been made frequently to me is that it would be a trivial matter to program a machine to answer "Yes" whenever the question "Are you certainly conscious?" was asked. Of course this is true, but such a machine would, by definition, not be reasoning according to the rules of *H*. Therefore such an objection has no relevance to the Theorem A.

It has also been suggested that the correctness of X does not have to be established for M to be certain that the evidence state K is correct, since in the possible world where X is correct, M may have access to other, different evidence from the possible world in which X fails.[14] However this argument implicitly assumes another evidence function, and Proof A will apply recursively to any such function. One can express this evidence function in terms of the question "Does this other, different evidence demonstrate the truth of some proposition?" Since Proof A is generalizable to all evidence functions, it will apply to this one. This is true regardless of what evidence M may have access to, since evaluating any such evidence is equivalent to the question "Does this evidence demonstrate the truth of some proposition?" and the function of answering such a question will always be subject to Proof A. No matter what evidence M may have access to, therefore, M can never be certain that the evidence function implied by that evidence has been correctly performed.

Some philosophers insist that any notion of certainty entailing an infinite sequence of propositions is epistemologically unacceptable and unnecessary, and that certainty requires only that M is certain of some proposition p. This would be a reasonable remark if the notion of certainty employed here were axiomatic, but for physical systems reasoning in *H* this is not the case. The requirement for an infinite sequence of functions is in fact a therorem in *H,* the proof of which is implicit in Proof A.[15]

[14] I am indebted to David Chalmers (personal communication) for this objection, and also for the one in the subsequent paragraph.

[15] I shall refer to this theorem as the *certainty lemma* in future work.

This proof can be expressed explicitly as follows: we start by assuming that M is certain of some proposition p. By the Principle of Physicalism, the process of determining p must supervene on some physical process, say X. From the Axiom of Fallibility, M can deduce that X may be fallible. By Assumption 1, M can deduce that if X is fallible, then the evidence state for p may be incorrect, in which case M can deduce that M cannot be certain of p. Therefore to be certain of p, M must have the ability to determine that X is operating correctly. By the Principle of Physicalism, this ability must supervene on some physical process, say X*. But what process is X*? It cannot be X, since M relies on the process X* to ensure that X is operating correctly. X therefore depends on X*, not the other way round.

Therefore M must assume that X* is some new process X', which is different from X. But from the Axiom of Fallibility, M can deduce that X' may be fallible. Since M's certainty of p depends on X, and M's certainty of X depends on X', M must have some means of determining that X' is operating correctly. Let us call this process X**. But what process is X**? It cannot be X', since establishing the correctness of X' depends on the correctness of X**. Neither can it be X, since establishing the correctness of X depends on the correctness of X'. Therefore we require some new process X''... and so on ad infinitum. This leaves us with no choice but to discard the assumption which led to the regress — that is, the assumption that M is certain of p.

A rather more serious objection to the proof of Theorem A is that the rules of *H* by which the proof is derived are never explicitly defined. The proof is derived using the rules of arithmetic and classical logic, and we must therefore assume, firstly, that our ability as human beings to use these rules is sound, and secondly, that our judgement that these rules are the correct ones to apply in this situation is also sound. This second judgement may be impossible to formalize. Somehow, *H* must allow us (and any system which reasons like us) to decide that it is classical logic, rather than, say, some paraconsistent or quantum logic, which is the correct logic to use in this case. The problem here is that it is impossible to say whether or not *H* is fully axiomatizable. If *H* can be fully axiomatized, then any valid informal proof in *H* will also have a valid formal derivation in *H*. The significance of this is that mathematical results are usually proved informally, to a standard of rigor which is demanding but nonetheless subjective. It is assumed that any valid informal proof will have a valid formal derivation in some axiomatized system, and this derivation can in principle be found by some purely mechanical process.

There are two difficulties which arise here. The first is that even if *H* is fully axiomatizable, we do not know what the correct axiomatization of *H* actually is. There is therefore no way of generating a formal derivation for any valid informal proof in *H*. We can however say that if *H* is fully axiomatizable, *there will definitely exist* some formal definition for any valid informal proof in *H*. It does not matter that we cannot say what this derivation is. If we can be satisfied with the validity of the informal proof, we can guarantee that a formal derivation of that proof is possible (provided *H* is fully axiomatizable).

It is in fact extremely rare in mathematics to require the derivation of a proof to be specified as a detailed derivation within some fully axiomatized system. Our ability as human beings to understand arithmetic, for example, also depends on the properties of *H*. It is similarly impossible to say that our ability to do arithmetic can be fully axiomatized, or, if it can, what that axiomatization is; but this does not prevent us from doing arithmetic, or from using arithmetic in mathematical proofs. One way of understanding this notion of an informal proof a little more clearly is to express it in terms of *effective procedure*. In computational theory, an effective procedure is any procedure which a human mathematician can use to compute some function on the natural numbers entirely mechanically — that is to say, by following a finite sequence of instructions by rote. An effective procedure requires only a pencil and an unlimited supply of paper, and no deep mathematical intuition or understanding.

The notion of an effective procedure is impossible to formalize. It is impossible to be sure that what appears to be a mechanical procedure to human beings is really a mechanical procedure in a deeper sense. For example, consider the problem of counting the sheep in a field. There is an obvious effective procedure for doing this — simply count the sheep one at a time until one arrives at the total. This is a purely mechanical, rote procedure for human beings, which leads one to suspect that it could be readily automated. Unfortunately, to count sheep, one first has to be able to recognize a sheep. This is a trivial problem for a human being, but a decidedly non-trivial problem for any automated pattern-recognition system. What appears on the surface to be a simply mechanical procedure, therefore, turns out not to be so simply mechanical on deeper investigation.

Despite this difficulty, there exists a widespread assumption in computational theory that all effective procedures are, indeed, capable of being automated. This assumption is known as the Church–Turing thesis. Here is one common formulation of this thesis:

Any effective procedure is Turing computable.

A procedure is *Turing computable* if it can be computed by some Turing machine. But we can broaden the concept of an effective procedure to include methods for proving conclusions using classical logic. An effective procedure for proving some theorem can therefore be defined as any method which allows a human being to prove that theorem by rote following of instructions, without requiring deep mathematical understanding or insight.

Since Turing machines can be regarded as equivalent to formal systems, we can express the Church–Turing thesis with respect to these sorts of effective procedures as follows:

Any effective procedure for proving some result has a derivation in an appropriate formal system.

One should note in passing that any effective procedure for proving some result must be a valid proof of that result. If we now assume that the problem of which formal system is appropriate in each case is itself effectively decidable, we can state the following principle, which we can call the Modified Church–Turing thesis or MCT:

Any effective deductive system is fully axiomatizable.

It is probably safe to say that most mathematicians do not worry about this very much, but to the extent that they worry about it at all, they implicitly assume the MCT is correct. If the MCT is correct, it still does not follow that *H* will be fully axiomatizable, since *H* may contain proofs that are not effectively describable. But even if we overlook this, the MCT seems to me to involve a considerable leap of faith. While it may be true, I for one am not prepared to take it on trust. Therefore I have neither assumed nor denied the MCT in proving the Theorem A. Readers should be aware, however, that the assumption that any valid informal proof can be formalized is in fact a version of the Church–Turing thesis, and this thesis is intrinsically unprovable.

A second difficulty arises if there is in fact *no* axiomatization of *H*. In this case, clearly, there might be no formal derivation for any proof in *H*. This extreme situation is in fact not very plausible — a more likely state of affairs would be if there were multiple axiomatizations which were contextual properties of *H*.[16] In this case there would exist separate formal systems each capable of proving certain sets of theorems provable in *H*, but no single formal system which could prove *every* theorem provable in *H*. In other words, the problem of which formal system to use in any given case would be formally undecidable. (It might still be *effectively* decidable, or perhaps decidable by means which cannot reasonably be described as effective.)

The conjecture that *H* is not fully axiomatizable is now known as the Lucas–Penrose thesis (Lucas, 1961; Penrose, 1989, 1994), and examination of this thesis is beyond the scope of this paper. The obvious consequence of this thesis is that it would entail some completely new way of understanding what it means for an

[16] A contextual property is a property whose value depends on the context in which it is observed or measured.

informal proof to be valid — whether this means that some effective proofs have no formal derivation, or that some results provable in *H* are not proved by effective procedures. In my proof of Theorem A I have neither assumed nor denied the Lucas–Penrose thesis; but it seems apparent that if human beings are indeed capable of self-certainty, they cannot perform this operation by means of any effective procedure. It is difficult to avoid the conclusion, therefore, that self-certainty in human beings would necessarily entail the Lucas–Penrose thesis.

Let us now turn to the question of what, specifically, constitutes humanlike reasoning. Examination of both Proof A and the proof of the certainty lemma above shows that in order to derive either proof a deductive system will need to include the following:

1. A capacity for *epistemic modal* reasoning, which is to say reasoning about the concepts of possibility and certainty;

2. A capacity for reasoning using classical logic;

3. A capacity for reasoning arithmetically.

Both Proof A and the certainty lemma involve primarily classical modal reasoning, but the final step involving the deduction of a non-terminating sequence requires arithmetic induction. A fully formal derivation of either Proof A or the certainty lemma would, however, require some means of encoding concepts such as "objectively real" and "physical existence." In the absence of any guarantee that such formal encodings are possible, I have opted for the simpler, rough-and-ready stipulation that a system capable of humanlike reasoning should be capable of following any procedure which human beings would regard as an effective procedure or mechanical algorithm. For a more detailed examination of the difficulties involved in relating informal proofs to formal proofs see Tanswell (2015).

A brief mention must be made of two alternative philosophical approaches which some people appear to believe constitute loopholes in Proof A. The first is *coherentism* — the doctrine that although individual processes may not be reliable, a large ensemble of processes together might be. This is obviously vulnerable to applying Proof A simultaneously to every process in the ensemble. Alternatively, the ensemble can simply be treated as a single process. More fundamentally, however, coherentism cannot be proven to be true; it must therefore be assumed to be true, which entails adding another axiom to *H*. Since the rules of *H* must allow M to ask the question "Is there any reason why this axiom should not be dropped?" and since the answer to this question is by definition NO, coherentism cannot establish absolute certainty.

The second philosophical approach which requires some mention is *constitutivism* (Shoemaker, 1990). The key principle here is that the evidence state for the question “Am I certainly conscious?” is a physical state which overlaps wholly or partly with the physical basis of consciousness itself. This idea was dealt with briefly in Reason (2016); the problem here is that, while self-certainty might be allowable if constitutivism is true, the process of determining if constitutivism is true can itself be represented as an evidence function, to which Proof A will of course apply. More specifically, for constitutivism to work, M must (by the rules of *H*) be able to answer YES to the question “Is the evidence state for consciousness physically identical with consciousness itself?” This question clearly implies an evidence function, to which Proof A applies.

References

Bell, J. S. (1964). On the Einstein–Podolsky–Rosen paradox. *Physics*, *1,* 195–200.

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

Chalmers, D. J. (2012). *Constructing the world*. Oxford: Oxford University Press.

Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research*, *45*, 153–176.

Kochen, S., and Specker, E. P. (1967). On the problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics, 17,* 59–87.

Kripke, S. (1963). Semantical analysis of modal logic I. Normal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, *9* (5–6), 67–96.

Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy, 36*, 112–127.

Parsons, T. (1996). What is an argument? *Journal of Philosophy, 93*(4), 164–185. http://philosophy.ucla.edu/wp-content/uploads/2016/10/WhatIsArg.pdf

Penrose, R. (1989). *The Emperor's new mind*. Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the mind*. Oxford: Oxford University Press.

Reason, C. M. (2016). Consciousness is not a physically provable property. *Journal of Mind and Behavior, 37*(1), 31–46.

Shoemaker, S. (1990). First-person access, *Philosophical Perspectives, 4*, 187–214.

Tanswell, F. (2015). A problem with the dependence of informal proofs on formal proofs. *Philosophia Mathematica, 23*(3), 295–310.