On Philosophically Undecidable Propositions of Philosophy of mind and Related Systems

Sam Hopkins

March 28, 2011

Contents

Table of Contents **1**

1 Preliminaries **3**

1.1 Introduction 3

1.2 Argument Synopsis 3

2 Mechanism **4**

2.1 First Formulation 4

2.2 Weakness and Strength Conditions on M 5

2.3 Caching out "Could" 6

2.4 Is the Mind Intrinsically Finitistic? 8

3 Mechanism and Gödel's Theorems **9**

3.1 Introduction 9

3.2 Benacerraf's Argument 10

3.3 Problems with Benacerraf's Argument 12

3.4 Penrose's Argument 13

3.5 The Final Argument 14
3.6 Aside: Regarding S and S^* 15
3.7 Conclusions 15

4 Half-Undecidability of Mechanism 16
4.1 Introduction 16
4.2 Holding that a Particular TM is Mechanistic 17
4.3 Believing M on Theoretical Grounds 18

5 Functionalism Implies Mechanism 19
5.1 Functionalism 19
5.2 The Basic Argument 19
5.3 The Case of Interactionist Dualism 21
5.4 Computability of Physical Laws 22

6 Conclusion 23

7 Appendices 24
7.1 Understanding Gödel's Theorems 24
7.2 Logic and Turing Machines 26

References 27

1 Preliminaries

1.1 Introduction

The announcement and proof of Gödel's two celebrated incompleteness theorems shocked the mathematical world in 1931. It seemed, if just for a few moments of intellectual history, that mathematics as we knew it had been crippled. The ensuing 80 years, however, have failed to carry out such dire predictions. Mathematics has more or less proceeded unhindered. Furthermore, attempts to ascertain what consequences (if any) the theorems have outside of mathematics proper have not been particularly successful: indeed, at this point, there is a good deal of eye rolling at such attempts in the philosophical literature [Sha98].

Nonetheless, this paper will be just that: an attempt to milk incompleteness for some philosophical consequences. So far as I know, the main argument presented herein is not present anywhere else in the literature, although I admit that I have not been as diligent as I might have in surveying such literature completely.

I am loathe to state too early what the main claim of this paper will be because, though it can be stated rather elegantly, it will take a great deal of discussion to elucidate completely what is meant by the claim. With the proviso that nearly every word it contains needs its meaning cached out correctly, here is the thesis:

> If functionalism is an adequate account of the mind, then we cannot know this to be the case.

1.2 Argument Synopsis

The argument will require a good deal of background discussion and build-up before the punch line can be delivered. I will formulate a useful intermediary thesis (one which ultimately is uninteresting in-and-of itself) which is essentially a precise "answer" to the age-old question of Mechanism. Thesis M

will say, roughly, that there exists a Turing Machine (henceforth TM) which enumerates the mathematical output of a person's mind, idealized in a certain way. Then I will subject the reader to a substantial amount of technical argument, the upshot of which will be a certain undecidability result about M. Furthermore, I will separately argue that standard versions of functionalism imply M. Then the argument will go like this (again with the proviso that all the terms are as-of-yet ill-defined):

1. If functionalism is an adequate account of the mind, then M.
2. But if M then we cannot know that M.
3. Therefore, if functionalism is an adequate account of the mind then we cannot know that functionalism is an adequate account of the mind.

2 Mechanism

2.1 First Formulation

I mean to attain some sort of undecidability result about functionalism. The backbone of the argument—the real work—will be done by Gödel's incompleteness theorems. But in order to begin to contemplate the consequences of the incompleteness theorems for philosophy of mind, we need some reasonable way to apply them. The problem, of course, is that minds are, well, minds, and numbers are, well, numbers. The bridge will be TMs, but we cannot leave it at that. We need a philosophically-legitimate way to link them up. As an aside, let me make clear here that I *will not* attempt to argue for or against such a mechanistic thesis.

The problem, of course, with saying something so vague as

> There exists a TM that simulates the mind.

is that it is entirely unclear what such a TM might look like. People receive input in the form of sensory impressions; TM's receive input in the form of

coded tapes, and the analogue holds for output. Hence, we will need more specificity. The standard move here is to make what prima facie appear to be substantial concessions in favor of the TM and employ something like the following thesis [Sha98]:

> (M) For every person S, there exists a TM that enumerates[1] all the mathematical truths that S could ever prove.

2.2 Weakness and Strength Conditions on M

Since this move is rather far from intuitive, let me pause here to note a few desiderata about the mechanist thesis. As a thesis to be argued for or against it is probably relatively useless. So many idealizations and concessions are required even to make much sense of it, let alone make sense of a stronger thesis about simulation not restricted to merely mathematical output, that in the end its truth or falsehood are unlikely to be of any great consequence to strictly philosophical matters [Sha98].

For the purposes of my argument, however, we will require a mechanist thesis satisfying two constraints. The mechanist thesis must be both

- sufficiently strong to yield the desired undecidability result *(the strength requirement)*
- sufficiently weak to be entailed by functionalism *(the weakness requirement)*.

The strength condition requires that we not make the thesis, say, "There exists a TM that enumerates all the mathematical truths that S has thus far proved," since such is obviously a finite set and so will not admit the intended Gödel result. Even "all the mathematical truths that S will in a lifetime

[1] This is a formal term: a TM M enumerates a (countable) set S if it "runs through" the contents of S—that is, if M is started on a blank tape, for every $n \in S$ there is a step s_n in the computation where to the left of the tape head will be all and only the elements $1 \cdots n$ of S (delineated in some appropriate fashion).

prove" is insufficient, for the same reason. The thesis must say something about the existence of a TM enumerating an infinite set, but since actualized human behavior is finite, we will need to say something more about this notion of "could ever prove" (or whatever surrogate notion turns out to work best).

It is due to the weakness condition that we don't want something like "There exists a TM whose machine states correspond one-to-one with the mental states of S". A detailed argument may be found in [Hop], but the basic idea is that TMs simply do not have enough structural richness to capture all of the types of causal relations that functionalism must allow to hold between mental states. It is also due to the weakness requirement that we have weakened the thesis by passing to mathematical truths: they are more readily formalized in a language understandable by TMs—without a thesis we can make sense of, we certainly cannot have a thesis entailed by functionalism.

2.3 Caching out "Could"

So, returning to the concern about finitude, we need to pull some sort of philosophical trick to get a thesis that makes a claim of identity between the output of some TM and an infinite set of mind-produced mathematical propositions. I do believe that the modal option employed above—"the set S_p of mathematical truths that S *could* prove"—is the right way to go, but we need to cache out "could" appropriately.

A first impulse is to cache it out with a possible-world semantics for modality, but that will not do here: such a formulation will not satisfy the weakness condition. In the various possible worlds in which S produces her different outputs she also may receive different inputs. Unless her mind "funnels" that infinite input space into a finite one, there is no way that a TM can account for all of her behavior on infinite possible input, for the mere reason of being, I should think, unable to formulate that input. The TM

receives no tape input, so all its input must be hard-coded into the machine. In any case, retreating to some intuitions about what a plausible mechanistic thesis might say, it seems very strange to assert that a TM could come up with all of S's possible behavior without having anything but finitely-much of S's input.

Instead, I propose that we cache out "could" as follows. The problem with speaking of the output of a single mathematician over her lifetime is that we lose the infinite character needed to satisfy the strength condition. But what if we just let that mathematician "run forever"? The problem with real-world mathematicians is, of course, that they peter out—they lose their creativity, or their ability to make new neural connections, or whatever it is. But a mind, considered in the abstract, ought to have no such issues. Of course, without great care, we will once again violate the weakness condition via the infinite input problem.

How can this be made precise? The idea, which I leave to better metaphysicians than I to sort out, is that multiple realizability (which is surely a correct thesis, whether or not you accept a fully-blown functionalism) implies that the same minds that our human mathematicians have could be instantiated in physical media that undergo little (or no) physical degradation. We can then (almost) take the union of all the truths produced by S's mind instantiated in every possible instantiating medium until such decays to the point of no longer instantiating S's mind to be the idealized mathematical output of S. But once again we have the infinite input problem to deal with, since the union of all the inputs received by all these instantiations could also be infinite.

The modification we want is this: all the instantiations that last beyond a certain point (a normal human lifespan, for example), must exist in the same or very similar input context—we might say the same mathematical community. That is, at time t (relative, say, to the birth of S), if t is greater than some fixed t_0 then every instantiation of S will receive the same set

of inputs. Furthermore, this set of inputs must be computable from a finite amount of information—some initial conditions, if you will. Now, it would not do to assume here that the behavior of all or even some finite set of the others around S will the computable given a set of initial conditions, as this would be to beg the question, but this is exactly what is needed: S's mathematical community needs to be simulated. Fortunately, we will see later that functionalism implies exactly this result, which will make M, cached out as thus, satisfy both the weakness and strength conditions in the conditions we care about: functionalism assumed true. We defer the arguments, which make up the rest of this paper. First we must quickly dispatch an objection:

2.4 Is the Mind Intrinsically Finitistic?

It might be objected here that "the mind doesn't work like that." That is, it might be objected that the sort of finitism that I'm trying to avoid is an essential property of minds: regardless of the physical media in which they are instantiated, minds "wind down" after a period of time. After all, certainly minds must change (mature, grow, etc.) as a result simply of operating for a while in some intput/output context. Perhaps by their very nature after a while all this change results in degeneration.

I simply see no reason to believe this, but giving a full discussion will take us too far afield. A full answer would require saying something about the way minds change through time. Does the functional description change? Or might we interpret the effects of aging as created by additional input parameters to the functional network? In some sense, though, it is not a matter that can be settled philosophically. Eventually either the psychologists or the neurobiologists will discover the cure for aging; if it is the psychologists then I'm wrong, the neurobiologists and I'm right. Since the thesis is plausible enough, I see no reason not to accept it pro tempore. Let us therefore move on, supposing that we have made a formulation of mechanism that makes

a claim about the existence of a TM which enumerates the infinite set of mathematical truths produced by a suitably-idealized person.

3 Mechanism and Gödel's Theorems

3.1 Introduction

Now I have some technical arguing to do. Adapting a canonical argument by Benacerraf, I will argue here for the thesis that if M then we cannot prove, for any TM T, that T is the machine guaranteed by M. [Ben67] proceeds in a highly formalized fashion; I will attempt here to limit the formalism while retaining all the necessary precision.[2]

At first glance, it would appear that the Benacerraf argument could just be used directly. After all, his purported conclusion is this:

> At best, Gödel's theorems imply... that given any Turing machine W_j, either I cannot prove that W_j is adequate for arithmetic, or if I am a subset of W_j then I cannot prove that I can prove everything W_j can. [Ben67]

However, I am fairly certain that Benacerraf's argument has a major failing: it proves too much, due to a terribly faulty assumption. I will give Benacerraf's argument here, though, because we can tear it down and build it back up with a few modifications to do what we want. The following argument, then, is as accurate a synopsis of [Ben67] as I can compress into this space. The notation is somewhat adapted.

[2] I must also here acknowledge [Sha98] for the excellent clarity it brings and for some of the notation, which I've stolen.

3.2 Benacerraf's Argument

Let S be the set of sentences I could prove,[3] and let S^* be its closure under logical implication.[4] We shall suppose that S^* (and so S) is consistent and furthermore (here's the bad assumption!), that everything I prove is knowably true (i.e., $(\varphi \in S^* \rightarrow \varphi)$, and this sentence is itself in S^*). From this follows $Con(S^*) \in S^*$.[5] Now suppose that there exists a recursively enumerable set[6] K such that

1. It follows from the things that I can prove that $Q \subseteq K$ (i.e., $(Q \subseteq K) \in S^*$)[7]

2. It follows from the things that I can prove that $K \subseteq S^*$ (i.e., $(K \subseteq S^*) \in S^*$)

3. $S^* \subseteq K$

(Note: Benacerraf could have just assumed $Q \subseteq K$, $(Q \subseteq K) \in S^*$, $K = S^*$, and $(K = S^*) \in S^*$. These more obviously correspond to the assumption that I am a Turing Machine and I know it (i.e. can prove it) But Benacerraf is trying to use the weakest assumptions he can, so we get these confusing ones instead. The latter set of assumptions clearly implies the ones Benacerraf makes, though, so any contradiction derived from Benacerraf's assumptions also follows from the stronger assumptions). Together these amount to the

[3]We will be careless and confuse the distinction between sentences and their Gödel numbers as much as we like.

[4]There is of course no reason to believe that S is already closed under logical implication: perhaps some proofs require too large a stroke of genius. Certainly, if I operate by mechanical derivation then it will be closed under logical consequence, but I don't, so it won't.

[5]In fairness to Benacerraf, he is responding to a badly-misguided paper by J.R. Lucas ([Luc61]) and so is working with this assumption I suspect only because Lucas does. In fact, in an appendix Benacerraf himself constructs an argument against the assumption, although he claims to believe that his argument can be overcome.

[6]A set is recursively enumerable if it is the codomain of a recursive function, or, equivalently, if there there exists an enumerating TM for it.

[7]Where Q is some minimal arithmetic; see, for example, [BBJ07].

assumption that there exists some TM T_K enumerating K such that (it follows from what) I can prove that T_K proves Q, (it follows from what) I can prove that I can prove anything that T_K proves, and, since we've assumed that I (knowably) only prove true things, $S^* = K$. Thus, T_K is a mechanistic TM.

By first incompleteness there is a Gödel sentence G_K for K. By second incompleteness, $(\mathrm{Con}(K) \rightarrow G_K) \in K$. Benacerraf here does something that I fail to understand. He distinguishes between this predicate, Con, and the one he used earlier: Con. He writes:

> Note that 'Con' in the antecedent of the sentence... is not italicized, while it is italicized wherever it appears previously in this proof. ... This is because ['Con(K)'] must be in the form appropriate to mirror the first half of [first incompleteness] in [K] itself. It is an abbreviation for some sentence in which [T_K] is coded in the proof predicate. In particular, we are not entitled to assume without proof that if [$Con(K) \in S^*$] then... [$\mathrm{Con}(K) \in S^*$]. We will therefore keep them at least typographically distinct. [Ben67] (changes are to preserve notation).

It is clear enough what is meant by Con, but Con is now a bit of a mystery. If we have not assumed that S^* is recursively enumerable in the first place, then there is no guarantee, at least from standard Gödel numbering, that the predicate involved in $Con(S)$ is even expressible in (first-order) arithmetic! So what our assumption that $S^* \subseteq K$ now means is rather unclear. Let us gloss over this concern for now, however, for there will be a knockdown (related) one.

Continuing with the argument: It follows from our assumptions and $Con(S^*) \in S^*$ that $Con(K) \in S^*$ (here the italicized, presumably informal consistency predicate), and it follows from second incompleteness that if a set X extends Q then the formal consistency of X is expressible in the language of X and is true iff the informal consistency sentence for X (i.e. $Con(X)$)

is. Furthermore, S^* knows about second incompleteness, so $\mathrm{Con}(K) \in S^*$. But then $\mathrm{Con}(K) \in K$, so K is inconsistent, so S^* is inconsistent, contrary to hypothesis. Thus, we must throw out one of the assumptions about K: there is no TM such that all of them obtain.

This is Benacerraf's argument. It is well-crafted and (mostly) reasonable (except for this $Con/\,\mathrm{Con}$ business). But it proves too much.

3.3 Problems with Benacerraf's Argument

The issue with Benacerraf's argument is that it shows too much: from Benacerraf's assumptions it follows very quickly that mechanism is entirely false. The issue is this assumption that $Con(S^*) \in S^*$. Suppose that you are presented with a TM J. Then you can of course effectively*i.e. in an algorithmic fashion* formulate the sentence $\mathrm{Con}(J)$. Furthermore, if $S^* = J$, then you know $Con(S^*) \equiv \mathrm{Con}(S^*)$, so $\mathrm{Con}(J) \in S^*$. But by second incompleteness this is impossible, so you could not both be a TM and *ever be presented with a description of that TM*. Surely it is not the case that I could ever be presented with a TM so complex that I couldn't formulate $\mathrm{Con}(J)$—it is an entirely algorithmic process! But the notion that a mechanistic TM for me might exist but I could never presented with it is absurd. So there is no such TM.

This is a very strong and I think rather implausible result. At any rate, we needn't rest on this implausible assumption that $(\varphi \in S^* \to \varphi)$ and $((\varphi \in S^* \to \varphi) \in S^*)$ for all φ. We can get the desired conclusion without the undesired one if we get rid of the hypothesis. All Benacerraf needs is $\mathrm{Con}(K) \in S^*$. In §3.3 of *Shadows of the Mind* ([Pen94]), Roger Penrose offers an argument that we can adapt to get this conclusion. We will not be able to argue conclusively that $\mathrm{Con}(K) \in S^*$ (as Penrose says, $\mathrm{Con}(K) \notin S$ will be a "bare logical possibility"), but we will be able to conclude that $\mathrm{Con}(K) \notin S$ is deeply implausible.

3.4 Penrose's Argument

The following is an adaptation of Penrose's argument. Mistakes are of course mine. Recall that K must be axiomatizible. Now, the rules of derivation for K can always of course be reduced to simply modus ponens by adding sufficiently-many axioms (the number will always be finite, because so is the set of rules of derivation); call this new system K'. Suppose first that the list of axioms for K is finite. Then so is the list of axioms for K'. Now, since $K \subseteq S^*$, there will come a time when all of the axioms of K' are proven, so since the only rule of derivation is modus ponens, $\mathrm{Con}(K) \in S^*$.

Now, suppose the list of axioms for K is infinite. Then the preceding won't work. Now, the list of axioms can always be reduced to a finite list, but we have to add rules of derivation in order to do so. Let K' be K with the axiom set made finite but (finitely-many) rules of derivation added. There is no assumption, recall, that the rules of derivation are somehow contained in S^*: just the contents of K itself. So, notes Penrose, every element of K is:

> (unassailably) acceptable, but [its finite set of rules of derivation] R contains at least one operation that is regarded as fundamentally dubious. All the theorems of $[K]$ would have to turn out, individually, to be things that can be perceived as true—somewhat miraculously, since many of them would be obtained by the use of the dubious rules of R. Now, although each of these theorems can *individually* be perceived as true (in principle) by human mathematicians, there is no *uniform* way of doing this. [Pen94] (changes are to preserve notation)

Why is there no uniform way of doing this? If the method were uniform, then it could simply replace the dubious rule in the rules of derivation. So, in order for us to prove every proposition in K (which, by hypothesis we can do), we must draw on an *infinite* set of proof rules, each of which must be so different from the other that they cannot together be expressed by

the addition of only finitely-many axioms to the system. Furthermore, as Penrose notes, the success of the dubious rule in each case (for of course by use of the dubious rule one could derive each statement in question) would appear rather miraculous.

By itself, this is by no means entirely convincing: I see no reason why we might not, at regular intervals proceeding indefinitely into the future, discover new methods of proof of whose validity which we are entirely convinced. If Penrose is right, then this argument may have stronger consequences than mere undecidability. At any rate, we can do better.

3.5 The Final Argument

Here is the final version of our argument. Let S and S^* be defined as before, but do not assume that everything in S (and so in S^*) is knowably true. Suppose that there exists a recursively enumerable set K with Benacerraf's three properties (as listed above).

Again by first incompleteness, there is a Gödel sentence G_K for K. Again by second incompleteness, $(\mathrm{Con}(K) \rightarrow G_K) \in K$. Therefore $(\mathrm{Con}(K) \rightarrow G_K) \in S^*$. To get to our conclusion, we need $\mathrm{Con}(K) \in S^*$. By Penrose's argument, if $\mathrm{Con}(K) \notin S^*$, then the list of axioms for K is infinite, and in order to prove every element of K we must draw on an infinite set of proof rules (since the finite set of proof-rules used to generate K will contain a dubious rule). But examine assumption (2). Surely, if I can prove that I can prove every element of K, then I can prove the consistency of K! If I had to prove that I could prove every element of K, then either I actually did so (i.e., the list of axioms was finite), or I somehow came to grasp the infinitely-many proof techniques. How I might do so is rather unclear, but what is fairly clear is that I could not have done so without being able to prove $\mathrm{Con}(K)$.

So, $\mathrm{Con}(K) \in S^*$. This leaves us with two options: either we have a contradiction or K (and so S^*) is inconsistent. I will not address the

latter possibility in any detail because I more or less dogmatically refuse to believe that it is possible. (See [Pen94] for a discussion.) However, I must address one objection at this point: I said earlier that Benacerraf's assumption of $Con(S^*) \in S^*$ was faulty, but I appear just to have assumed it myself (dogmatically, at that!). In a sense I have, but we might say that I have assumed it philosophically rather than mathematically. It is a necessary assumption to get off the ground philosophically, but that does not mean we are sufficiently confident in it, in arithmetized version, to place it into the body of mathematical truths for which we suppose we have proofs. That is, I see no *mathematical* reason that $Con(S^*) \in S^*$, but things can only be in S^* for mathematical reasons.

3.6 Aside: Regarding S and S^*

It may have become apparent to the astute reader that I have been a little careless in distinguishing between S and S^*. My argument has concluded (roughly) that there is no recursively enumerable set K such that $K = S^*$ and it is a logical consequence of what I could prove that $K \subseteq S^*$. I really want a conclusion about K and S, though. But this is easy to attain: the reader can check that if there is such a TM with respect to S then there will trivially be one with respect to S^*, so the conclusion follows by modus tollens.

3.7 Conclusions

I will not offer any further argument in this direction, but I will take it that we have shown that there is no TM such that I could prove of that TM that its output is a subset of mine and that TM's output in fact matches mine exactly. We now move on to somewhat less technical arguments.[8]

[8] Astute readers may have noted that I have not considered the possibility that the first of Benacerraf's assumptions is the one to be thrown out. This is because I think throwing such out is highly implausible. By assumption (3) and the fact that I can prove Q (or

4 Half-Undecidability of Mechanism

4.1 Introduction

Our question for this section is as follows: could one hold mechanism to be true without being able, *in principle*, to exhibit a mechanistic TM? If I can convincingly argue that such is implausible, then we will have a half-undecidablity result about mechanism: if mechanism is true then we couldn't know it. I will make an attempt at such an argument. The idea will be this: any good reason for holding that M will have behind it an argument that could in principle be refined to exhibit a mechanistic TM along with (the right sort of) a guarantee that the exhibited TM is in fact mechanistic. As above, this is impossible (as long as the guarantee is mathematizable).[9]

A first point is that there is a strong sense in which M ceases to be verifiable, in that if true it cannot be directly checked. For how could it be directly checked? We would hold up a TM and say "this is the mechanistic TM!" But we've established that such is not an option, at least if that statement is to be made with the certainty of mathematical proof.

Might someone make such an assertion without the certainty of proof? Well, we must ask first what a "proof" might look like (an issue I have conveniently side-stepped until now). I must admit to not having a particularly complete account. The obvious issue is that we do not usually "prove" things about people but rather only about mathematical structures, and regardless of whether there is a TM that simulates you, you are by no means a mathematical structure. But it is a pattern in the history of science that theories eventually become formalized mathematically. (and this is especially true

at least Q is among my axioms, depending on the favored philosophy of mathematics), $Q \subseteq K$. So it would have to be the case that I could not prove such, for (1) to be false. I do not find this particularly plausible: it would be very strange if I could prove that $K \subseteq S^*$ but not $Q \subseteq K$. At any rate, it doesn't matter much: the concerned reader should feel free to insert "either I cannot prove that $K \subseteq Q$ or..." in the appropriate places.

[9] I would like to thank Jonathan Ettel for characterizing the argument to me in this fashion.

if we believe that the special sciences reduce to physics, which most clearly among sciences admits mathematical formalization). But we *can* construct proofs about the structure that results from that mathematical formalization.

In these terms, the preceding formal result comes to this: if M is true, then the mathematical formalization of our "final" or "perfect" scientific theory of the brain/mind will not be such that the mathematical outputs it predicts will be provably equivalent to those of any TM.

4.2 Holding that a Particular TM is Mechanistic

I suppose the reasons someone might hold M at this point could fall into roughly two categories. First of all, someone might hold that some TM in particular was the mechanistic TM. This would of course be held without the certainty of proof, but perhaps someone could have less-strong but nonetheless-convincing reasons. It is somewhat difficult to fathom what such reasons would look like. It seems unlikely that they would be completely nonformal—we would laugh out of town anybody who said only "this TM is mechanistic—I can feel it!"

There could also be more formal reasons, short of proof. Perhaps we have been able to prove of some TM J that it proves every piece of mathematical output that we have thus far ourselves proved, and whenever we prove something new we check to see whether J also proves it and, sure enough, it does. Furthermore, when we simulate J (which may be a very slow enterprise: it is to be expected that the complexity of J is on the order of the complexity of the brain) and it generates some new output φ, we always find that with enough labor we are able to ourselves generate a proof for φ. Prima facie, this might be a good reason to believe M. But it is not: at any fixed time there are guaranteed to be TMs (infinitely-many TMs, in fact) for which these observations have been historically true. The existence of one therefore provides no evidence at all in favor of M.

The only other sort of formal reason short of proof that comes to mind

would be an almost-proof of input/output equivalence between our best mathematically-formalized model of the brain and some TM J.[10] By "almost-proof" I mean a solid argument with a hole. These sorts of things abound in mathematics: frequently, they are arguments that were taken for years to be proofs until some issue, special case, or even counterexample was found. Certainly, if M is true, then we could reasonably expect to generate such arguments. The question is whether or not we could expect to if M is not true.[11] Now, if we are in this epistemic situation then we know that M has not been decided, so we could legitimately propose of some TM J that it is mechanistic. It is doubtless the case that arguments would be proposed that J is mechanistic, and those arguments would necessarily have holes, in this case because M is false. But those holes could doubtless be further and further disguised until the argument appeared reasonable (although the error would be exposed with sufficient inspection). Thus, I see no reason why there might not be this sort of almost-proof even if M is false: the presence of almost-proofs also fails to provide a good reason to believe M.

4.3 Believing M on Theoretical Grounds

There is a second sort of reason that someone might hold M. Perhaps there is some other position that implies M, and that position is held sufficiently strongly that M is accepted as an unverifiable consequence. (We will see that functionalism could be such a position.) Of course, the implication could not be of a sort that offered a construction of a mechanistic TM, because we would run into all of the above problems. So, somehow, this position has to guarantee the existence of a mechanistic TM without offering a provably-correct candidate. Even more: it cannot offer a candidate *even in principle:*

[10] Where "input/output" equivalence is short for whatever sort of equivalence between the two structures turns out to be necessary for J to satisfy M, no more.

[11] Note, of course, that if M is not true then nothing I have said precludes us from knowing so: we are only interested in our epistemic situation should M in fact be true. But when considering ways our epistemic situation should M be true, we must examine what sorts of things might or might not count as reasons for or against M in general.

a theory that in principle offers a candidate but is too complex for that candidate to be practically extracted will also be precluded.[12] Without any particular theories here to inspect (except for functionalism and more general physicalism, which will be discussed in the next section), there is little more to say, other than to note the following: it would have to be a powerful position indeed to make us believe that there exists a fairly straightforward (to formulate, anyway) mathematical theorem that is completely beyond the powers of our proof.

5 Functionalism Implies Mechanism

5.1 Functionalism

I do not have very much new to say about functionalism, but I need to lay out our terms and basic definitions. Functionalism is normally presented as a thesis about mental states: mental states are loci (if you will) of causal roles. We, however, need functionalism to be an account of the mind as a whole. The translation is not difficult so long as we assume that the notion of instantiation here is not difficult. "Functionalism is an adequate account of the mind" means that if S and T instantiate the same network of mental states (functionally-defined) then they have the same mind.[13]

5.2 The Basic Argument

Suppose that functionalism is an adequate account of the mind; I will argue that the mechanistic thesis holds.

Now, the conclusion requires no argument at all if our functionalism is a true machine functionalism.[14] So the question becomes: is causal-role

[12] I don't have the space or skill to argue it, but I think the condition will actually be that the candidate cannot be Turing-computably extracted.

[13] Of course, the problem of identity of minds is a tricky one, but not more so here than anywhere else, so I do not propose to treat it.

[14] By "true" I mean that machines are not just used as an analogy.

functionalism equivalent to a strong-enough machine functionalism?

Suppose, for the minute, that we are not interactionist-dualists. A materialist-functionalist must believe that minds are wholly instantiable by physical stuff, and an epiphenomenalist-functionalist must believe that all the bits of the mind responsible for mathematical output are instantiable in physical stuff (by the definition of epiphenomenalism). Let us proceed for the moment as though we have no reason to believe that the laws governing our physical universe are not computable.

Mechanism follows quickly but not immediately. A first attempt: the mechanistic TM needs merely to start with an initial state of whatever bits of our mathematician S instantiate S's mind and get to computing. By not simulating any physical degradation, the mechanistic TM will by definition arrive at the set specified by our cashing-out of the mechanistic thesis. The issue, of course, is the problem of input that we have already encountered above: over the course of an "infinite" lifespan, S receives all sorts of input that the TM cannot have. It is totally unreasonable to presume that S's collection of produced mathematical truths would be the same if S had received substantially different (or indeed no) input than she did.

It is reasonable, however, to suppose that S will have the same produced-theorem set under minor variations in (especially non-mathematical) input.[15] But the mechanistic TM can simulate input at least arbitrarily close to the input that would be experienced by the various instantiations of S (the sum of which, remember, are infinite-S) simply by simulating all the various phys-

[15]Well, at least it is by my lights. It is less than obvious why, however, especially since in ordinary people we do expect minor variations in input to produce (minor) variations in output. Here we encounter the advantage of restricting ourselves to mathematical behavior. Far be it from me to tread to far into this field of infinitary psychology, but it strikes me as likely that a (competent) infinite mathematician will eventually discover all of the easy (i.e. routine) consequences of each previously-derived theorem, so all that really matters is the momentous inputs: the inputs that spark a deep new insight or idea. Thus passing to the infinite mathematician, as we must do anyway, has the advantage of discretizing the behavior of the output-set with respect to the input-set and so allowing some margin of error in the TM's simulated input.

ical processes nearby all the instantiations of S, including other brains which might be S's teachers, collaborators, etc..[16] This is possible—i.e. the requisite finitism is preserved—precisely because we have required that each instantiation be in the same (mathematical) input/output context after a certain point.

5.3 The Case of Interactionist Dualism

Suppose though that we *are* interactionist dualists. We must be a particularly strong-toothed breed of such, because we will have all sorts of bullets to bite. First of all, not only must we believe that in this possible world people have ghostly bits (or mental properties)—we have to believe that such is true in *all* possible worlds, for obvious reasons.

What does this strong-interactionist-functionalism look like? It will by definition say that functionalism is an adequate account of the mind, and it will say that the resulting causal network is uninstantiable in any physical system: only ectoplasm can instantiate it. Since there are possible worlds in which there are arbitrarily *complex* physical systems, the resulting causal network (i.e. mind) must employ a causal relation of an entirely different kind from those instantiated in the physical systems with which we are familiar. Furthermore, to propagate this causation relation indefinitely, as a TM is perfectly capable of doing for ordinary sorts of causation, must somehow be uncomputable! This, it must be said, is utterly strange, and rather quickly places the burden of proof on the interactionalist-functionalist. Since it is not the place of this paper to venture into the details of ectoplasmal causation, I will leave this exotic viewpoint behind (who in the world would want the commitments of functionalism if they've gone interactionist anyway?), I think satisfactorily defused by its own apparent strangeness.

[16] It may or may not be the case that the TM can in fact get the inputs *exactly* right—in any case I don't think I need that strong a claim here.

5.4 Computability of Physical Laws

Let me return to supposing that we are not interactionist-dualists. Above I have been forced to assume the computability of physical laws. This is not necessarily such a good assumption. There are various results suggesting that there may be fundamentally uncomputable physical processes. (That is, there are some equations that are thought to govern very many physical phenomena—the wave equation,[17] for example—for which there exist solutions whose values cannot be computed at various points in time.[18]) See, for example, [Kre82].

The astute reader may have noted that, until now, all of my arguments would have gone through for any old physicalism without requiring much modification. If physical laws might be uncomputable, though, the arguments fail. But I have, in fact, assumed that I am working with functionalism as a particular theory of mind, and this will save the argument from worries about uncomputability of physical laws. If functionalism is an adequate theory of the mind, then all that is required of a mechanistic TM is to simulate the functional network, not the bare physical phenomena.

Admittedly, without there being any particularly well-spelled-out versions of causation in relation to functionalism (or, at least, any that are known to this author), there is not all that much to say here. As above, in order for the simulation of the causal-network model implied by funcationalism to fail, the functionalist will have to admit to his causal network some sort of uncomputable causation relation. The TM model of computability is very robust (i.e. the class of computable functions is stable under a wide variety of modifications to the model), so this uncomputable causation relation will prevent minds from being instantiated by a very wide range of

[17] Not to be confused with Schrödinger's equation.

[18] I am of course glossing over here what it might mean for a TM to compute a function on the reals; presumably for a function to be computable on the reals just means that it is computable to arbitrary precision with a sophisticated-enough TM

automata.[19] This seems to me to go somewhat against the spirit of functionalism. Since the applicability to our world of these theoretical results regarding the computability of physical laws is still under debate, I will leave this issue somewhat open.[20]

6 Conclusion

Where have we come? I suppose there are a few ways my argument could be interpreted. First there is the very strong conclusion (the one I gave at the beginning of the paper): if functionalism is an adequate account of the mind then we couldn't know it. Though I do believe that this conclusion is true, skeptics may believe that various minor holes in my argument cannot be overcome. In that case, the conclusion might instead be considered to be this: if functionalism is an adequate account of the mind, then in order to avoid contradiction a functionalist must hold that, no matter how refined psychology should become, a provably-mechanistic TM cannot emerge, or there is some uncomputable causation somewhere in the network of mental states. This much weaker conclusion does seem to me to have substantial implications for psychology.

Furthermore, the argument has some implications even if we through functionalism out the window for a minute and just consider some sort of bare physicalism. Unless this bare physicalist also believes that for some reason the mechanistic TM whose existence is guaranteed could not *in principle* be

[19] For further discussion of what sorts of machines minds might be if not Turing Machines, see [Cop98].

[20] Astute readers may have noticed one further problem here. I noted way back in §2.3 that the argument here would make our caching-out of M satisfy both the weakness and the strength condition. Functionalism does indeed imply that the input S receives from the people around her can be simulated, but it does not necessarily imply that other environmental input can be so simulated—the chirping of birds, the appearance of trees, etc.. However, (although this is not a true response) the fact that modern video games can already generate such realistic environments on machines that are nowhere near as powerful as a true TM (i.e. one with an unbounded memory) suggests that this is not a particularly serious issue.

exhibited (a strange view indeed), he will have to countenance some sort of Penrose-ian uncomputability of physical systems. See [Pen94].

Finally, a little perspective. The whole debate here—the existence of the sets S and S^*, for example—relies on some fairly strange idealizations. Now, because I have constructed thesis M as merely a useful proposition, not something I mean to argue for or against, I may reasonably claim somewhat more license to invoke these idealizations than would someone actually arguing for or against a mechanistic thesis (this is how I might, provisionally, answer the objections to idealization in [Sha98]—I do not claim that M is in and of itself philosophically interesting). However, the idealizations I propose and my claim that they will result in M satisfying the weakness and strength conditions still have the flavor of a philosopher reaching too far outside his domain of expertise. Should science eventually bear out my idealizations to be plausible, then I do believe that with sufficient additional detail at very least the second sort of conclusion I propose will be valid. But we must remember that the weakest link in all of this arguing is not the central arguments themselves but rather their dependence on a thesis whose cogency depends on scientific results yet unproven.

7 Appendices

7.1 Understanding Gödel's Theorems

This appendix is a crash course on Gödel's Incompleteness Theorems. The exposition is inspired by [Ben67]. The reader is referred to, for example, [BBJ07] for definitions and proofs.

Gödel's brilliant idea was to make systems of logic to talk about themselves. The key move is to make formal languages into just so much arithmetic. Each symbol in the language is assigned a sequence of numerical digits in such a fashion that no two sentences map to the same number. Rules of derivation will correspond to arithmetical manipulations (add 5, multiply by

10, etc.). Additionally, we can code readily code sequences of such numbers themselves as numbers (by, for example, exponentiating primes) and so there will be a number-theoretic predicate $D(x, y)$ expressing "x is a derivation of y from the null set".[21] Furthermore, for any axiomatizable theory T, there will be a predicate $D_T(x, y)$ expressing "x is a derivation of y from the axioms of T".

Now, all that is needed to express this predicate is the language of number theory, so it is to be expected that some formal languages—indeed, exactly the ones we want to use to do mathematics—will be able to express D_T. More particularly, we restrict ourselves to theories which contain a minimal arithmetic (say, Q) as a subset; this will allow them to simulate derivation in the right ways. Fix such a theory T. Then, essentially by a sophisticated diagonalization, we can formulate a sentence G:

$$(\forall x)\neg D_T(x, \mathbf{G})$$

where $\mathbf{G}$ is the number of G. G is called a *Gödel sentence for* T, and it "says" that there is no derivation of G from the axioms of T. Now, because T contains all the tools necessary to simulate derivation, it will be the case that for any theorem $t \in T$, the sentence $(\exists x)D_T(x, \mathbf{t})$ is also a theorem.

Suppose $G \in T$. Then so is the sentence $(\exists x)D_T(x, \mathbf{G})$, which contradicts G, so T is inconsistent. So suppose $\neg G \in T$. Then the sentence

$$(\exists x)D_T(x, \mathbf{G})$$

is also in T. But since we have supposed $\neg G \in T$, for any number n it turns out that Q is strong enough that $\neg D(n, \mathbf{G}) \in T$ (i.e. T "knows" that no number codes a derivation of G). But then T is ω-inconsistent, since $\neg G$ says that there is such a number. Thus we have Gödel's first incompleteness

[21] More formally: there will be a sentence in the language of number theory $\varphi(x, y)$ with two free variables such that $\varphi(a, b)$ is true under the standard interpretation of arithmetic iff a is a valid derivation of b.

theorem:

Theorem (Gödel)**.** For any axiomatizable and ω-consistent theory T in the language of arithmetic such that $Q \subseteq T$, there exists a sentence G such that neither $G \in T$ nor $\neg G \in T$. Actually, of course, this is a weak form of the theorem. The result actually holds for any consistent theory (that is, we can replace ω-consistency with consistency), but the particular sentence G we have constructed is insufficient to prove it.

The second incompleteness theorem can be largely understood with just the material I have just presented. We typically define a theory T to be consistent iff it does not contain a sentence of the form $\varphi \wedge \neg\varphi$. Observe, though, that for theories containing Q, this is equivalent to the assertion that $(2 \neq 2) \notin T$, so we take that as the definition for present purposes. Then, of course, there is a sentence in the language of arithmetic $\mathrm{Con}(T)$ expressing that there is no derivation of $(2 \neq 2)$ from the axioms of T.

Theorem (Gödel)**.** For any axiomatizable consistent theory T in the language of arithmetic such that $Q \subseteq T$, $\mathrm{Con}(T) \notin T$.

7.2 Logic and Turing Machines

This appendix contains a proof sketch of the result that there exists a one-to-one correspondence between enumerators and axiomatizible theories, which is important for the technical section of this paper.

Let T be an axiomatizible theory. Then let M_T be a TM with the following behavior: M_T simulates an enumerator of $\mathbb{N}$. At each number n, M_T checks to see if n is a valid Gödel number for a derivation in our formal language. If it is, M_T locates all the assumptions in the derivation and checks whether or not they are in the set of axioms (which is by definition recursive). If they are, M_T prints the last line of the proof coded by n. If not, M_T moves on to $n + 1$. Therefore M_T enumerates T.

Let M enumerate some set S_M closed under logical implication. Then

there exists an axiomatizible theory $T = S_M$: just let the axioms of the theory be S_M![22]

References

[BBJ07] George Boolos, John Burgess, and Richard Jeffrey, *Computability and logic*, 5th ed., Cambridge University Press, 2007.

[Ben67] Paul Benacerraf, *God, the devil, and gödel*, The Monist **51** (1967), no. 1, 9–33.

[Cop98] B. Jack Copeland, *Turing's o-machines, searle, penrose and the brain*, Analysis **58** (1998), no. 2, 128–138.

[Hop] Sam Hopkins, *Could minds be (anything like) machines?*

[Kre82] Georg Kreisel, *Review [untitled]*, The Journal of Symbolic Logic **47** (1982), no. 4, 900–902.

[Luc61] J. R. Lucas, *Minds, machines and gdel*, Philosophy **36** (1961), no. 137, 112–127.

[Pen94] Roger Penrose, *Shadows of the mind*, Oxford University Press, 1994.

[Sha98] Stewart Shapiro, *Incompleteness, mechanism, and optimism*, The Bulletin of Symbolic Logic **4** (1998), no. 3, 273–302.

[22]Of course, this doesn't quite work: S_M is only recursively enumerable, not recursive. This turns out not to matter: see [BBJ07].