

# Paraphrase, Ontology, and the Index Theorem

Sharon Berry

May 11, 2010

## 1 Intro

In this mini-essay I will argue that a variant of the Index Theorem has an interesting philosophical consequence: it suggests an a priori reason why a priori we might never be able to discover/recognize suitable paraphrases of certain higher level sentences into the language of microphysics. I will describe how two radically different morals can be drawn from this fact, and then advocate a less controversial general conclusion.

## 2 A variant on the Index Theorem

Different computer programs can compute the same function, in the sense that they give the same output for all possible inputs (for these purposes let's say that the input to a program is a natural number). The question of whether two different programs are indeed computing the same function can be very very hard. In fact, the following little variant on the index theorem shows that if  $\psi$  is a program, no helper program is capable of recognizing all the other programs which compute the same function as  $\psi$ .

Suppose there were some helper program that recognized all programs that calculate the same function as some partial recursive function  $\psi$ . Then there would be a program that solves the halting problem as follows. Either  $\psi$  is undefined for some  $n$ , or not.

If  $\psi$  is undefined at some  $n$ , then we can determine whether the eth program halts on input  $x$ , by alternately running the eth program on input  $x$ , and the helper program applied to the function which emulates  $\psi$  for all values other than  $n$ , but on input  $n$  emulates the eth program when run on input  $x$ . If the eth program halts on input  $x$ , we know that it halts. If the eth program does not halt on  $x$  then the latter patched together function behaves like  $\psi$  on all values (because remember  $\psi$  doesn't halt on input  $n$ ) so the helper program will eventually recognize it, at which point we know that the eth program does not halt on  $x$ .

If  $\psi$  is defined on all  $n$ , then we get a solution to the halting problem as follows. Consider the program that, for each input  $n$ , calculates the first  $n$  steps of the eth program applied to  $x$ , and then goes into an infinite loop if

this program halts at that stage, and otherwise then just calculates  $\psi(n)$ . This program behaves the same as  $\psi$  if and only if the eth program applied to  $x$  never halts. Hence if alternate running the helper program on the code for this patched together program, and the eth program at  $x$ , we will again know whether the eth program applied to  $x$  halts. If it does halt, we will see that directly. If it doesn't halt, then our patched together program calculates the same function as  $\psi$ , so our helper function will eventually halt and output a positive verdict, from which we can conclude that the eth program applied to  $x$  does not halt.

### 3 What this suggests about the search for reductive paraphrases

Now, if we think of our own brains as computers, which take our contact with the sensory world as input and yield a verdict about whether some sentence is true or false (or a failure to come to any conclusion at all) as output, this suggests the following possibility. There may be many sentences which reasoning would, in fact, always ultimately assign the same truth value to in all possible situations. But we may never be able to figure out that these two different sentences are related in this way. For any partial recursive function  $F$  (the kind of function that is represented by the behavior or a turing machine which may or may not halt), and any computable helper function representing our mathematical insight, there will always be some possible other program which actually computes the same function as  $F$ , but which the helper function cannot recognize to compute the same function.

I claim that this uncontroversial mathematical fact has an important consequence for the prospect of physicalist reductions and ontology generally. The direct consequence is that physical reductions can be very hard. A physical and a non-physical sentence can be equivalent in the sense that our reasoning will always (eventually) lead us to exactly the same verdict about whether each sentence is true, and yet it might be impossible for us to ever recognize this fact. Even if some physical sentence "There are atoms behaving in such and such a way" and some non physical sentence "There is a liver" were evaluated by processes of reasoning which always eventually to exactly the same results, it might well still be the case that we could not \*recognize\* this relationship between the two sentences. Even if psychological investigation told us exactly what methods we used to evaluate claims about each of these sentences, and even if these methods were completely reliable, and we knew that they were reliable - even then we might still not be able to assure ourselves that the two methods always yielded the same verdict. Hence, (plausible) we could not know that a reductive definition which replaced the claim "There is a liver" by the claim "There are atoms behaving such and such a way" was correct. All experience could tell us would be that our intuitive methods for evaluating these cases had always yielded the same result in the cases we had so far considered.

It's a somewhat more complicated question what \*general\* conclusion to draw from this fact about the potential difficulty of recognizing correct physicalist reductions. At the very least, the considerations above raise serious questions about a popular program: evaluate whether physicalism is true by seeing if there are non-physical terms for which we cannot arrive at acceptable physical paraphrases. For, consideration of the index theorem suggests that we know in advance that there could be such terms.

However there are two - diametrically opposed - more general conclusions one might draw from this advance knowledge.

## 4 Two opposing philosophical conclusions

The physicalism-friendly alternative, is to say that consideration of this stuff about the index theorem shows the following: If we imagine a world that is as physical as possible, this world could well still \*appear\* to have lots of non-physical properties - in the sense that creatures in that world would not be able to \*discover\* which combinations of sentences about microphysical properties their high level sentences are equivalent to. They would have one pattern of reasoning about whether some sentence about microphysical properties M holds, and another pattern of reasoning which determined whether some higher level sentence H held, yet they still would not be able to realize that the situations in which they judge the higher level sentence H to be true are always and only exactly the ones on which they would judge the microphysical sentence M to hold. Hence, the fact that it turns out to be fiendishly difficult to figure out which (doubtless terribly complicated and disjunctive) combination of microphysical claims, a claim about livers is equivalent to, doesn't give any evidence that physicalism is false. All we are seeing at work here is the fact that it's a substantive (and in some cases impossible) task to recognize that two different methods of reasoning are really just different ways of thinking about the same microphysical state of affairs.

The anti-physicalist moral for this little story is just the opposite. According to it, what this thought experiment shows that physicalism is not only actually false, but it could never be true. Even if we try our very best to imagine a purely physical world, of atoms in the void, it will still turn out that there are claims of this world which (although identical in the situations under which they are accepted and would be true) are not in fact equivalent to the microphysical claims, claims like those which would be described by H sentences in the description given above.

## 5 A more general moral

I won't take a stand now on which of these responses is right with regard to physicalism. What I do want to do is draw a moral that's even more general than either of these two. The moral is this:

To the extent that physics gives us a certain picture of the world (e.g. as atoms in the void), this \*either\* does not give us reason to believe that not all objects are physical, \*or\* it does give us reason to expect that all objects are physical, but we should not further expect to (always) be able to discover which paraphrases which correctly replace apparent reference to non-physical objects with some combination of claims about physical ones. For, even the crudest physical picture of a world consisting of atoms in the void does not give us reason to think that the creatures in that world would be able to discover correct reductions of their higher-level sentences which don't appear to talk about atoms in the void to ones that do.