Gödel vs. Mechanism:

An examination of Three Attempts to Show that Gödel's Theorems are Incompatible with Mechanism and Why They Fail

Eamon Darnell

2013

Contents

Introduction

Kurt Gödel's incompleteness theorems have had a major impact in logic and philosophy of mathematics. It is not entirely obvious that the incompleteness theorems should be taken to have any significant implications outside of those two areas, although some have argued that they do. In particular, some have argued that Gödel's incompleteness theorems imply that Mechanism is false. Mechanism can be understood as the view that the human mind is a sort of computational machine which can be exactly modeled by some Turing machine or formal axiomatized theory (I will use the term formal system to refer to either of these). Gödel demonstrated that (certain types) of formal systems are limited with respect to what they can prove in arithmetic. If the human mind is not subject to these limitations, then it cannot be exactly modeled by a formal system and so Mechanism is false.

Below I will examine three attempts to show that Gödel's results are incompatible with Mechanism. The first, and perhaps best known, is developed by J. R. Lucas in his paper, "Minds, Machines and Gödel" (1961). Lucas argues that his understanding of Gödels results allows him to outstrip any particular formal system which could be proposed as an exact model of his mind. The second is developed by Roger Penrose in Shadows of the Mind (1996). Penrose argues that in light of Gödels theorems, the assumption that Mechanism is true leads to contradiction and so Mechanism must be false. The third is developed by Storrs McCall in his paper, "Can a Turing Machine Know that the Gödel Sentence is True?" (1999). McCall argues that Gödel's results reveal the existence of a category which humans can know about and use, but which formal

systems cannot know about or use. I will consider each of these arguments in turn and show that each fails to establish that Gödel's theorems imply that Mechanism is false. I will conclude, based on this and a weakness common to all three that it is improbable that Gödel's incompleteness theorems can be definitively shown to be incompatible with Mechanism.

1 Setting the Stage

Before examining the three arguments against Mechanism, I should first provide a brief overview of Gödel's results and of how Mechanism should be understood in the context of this discussion. This section is devoted to setting the stage for future discussion by providing a brief explanation of key concepts and outlining a key assumption. I have divided this section into three main subsections. In section (1.1), I provide a brief overview of Gödel's results. In section (1.2), I explain how Mechanism should be understood in the context of this discussion. In section (1.3), I describe an important assumption made by at least two of the authors whose arguments I will be examining.

1.1 Gödel's Incompleteness Theorems

To understand the discussion that will follow, it is important to know what Gödel's theorems say. Gödel's incompleteness theorems apply to any formal axiomatized theory T, where T is ω-consistent [1] and strong enough to capture

[1] A formal axiomatized theory T, is ω-consistent just in case it is not the case that for some open formula, $F(x)$ T proves $\exists x F(x)$, yet for each number, n, T proves $\neg F(n)$. It was later shown by J. B. Rosser that Gödel's result can be extended such that one need only assume mere consistency to show that sufficiently strong formal theories are incomplete. Many authors who use Gödel's theorems to show that Mechanism is false speak only in

a basic amount of arithmetic. Gödels incompleteness theorems yield three results which are of particular importance to the present discussion. The first is the first incompleteness theorem (G1). G1 states: *if T is ω-consistent, then T is negation-incomplete.* To demonstrate this, define a Gödel sentence G_T, for T, such that G_T (indirectly) says, "G_T is not provable in T." The definition of G_T is such that if T proves G_T, it proves $\neg G_T$ and if it proves $\neg G_T$, it proves G_T. [2] Hence, if T is ω-consistent, G_T must be undecidable in T. The second important result is that T can prove that if T is consistent, then G_T (is true). That is, T proves the sentence, $Con_T \rightarrow G_T$, where 'Con_T' is the (indirect) statement, "T is consistent". This is important for demonstrating the third important result: the second incompleteness theorem (G2). G2 states: *if T is consistent, T cannot prove its own consistency.*

1.2 Mechanism

For the purposes of this discussion, *Mechanism* should be understood as the view that the human mind is a sort of computational machine which can be exactly modeled by some formal system. More relevant to this discussion is that Mechanism holds that whatever arithmetical sentence one's mind can prove can also be proved by a formal system which exactly models one's mind. That is, for any person P, there is a formal system M, such that M proves all and only the arithmetic sentences that P proves. Put more succinctly, M

terms of mere consistency. It is perhaps worth noting that doing this requires the change from Gödel's original undecidable sentence for a theory T which indirectly says of itself, "I am not provable in T" to Rossers undecidable sentence which indirectly says of itself, "If I am provable in T, then T already proves my negation." In the later sections of this paper I speak in terms of mere consistency except where ω-consistency is especially relevant.

[2] Note: establishing this conditional requires the assumption that T is ω-consistent.

exactly models P just in case:

$$\forall\psi((P \vdash \psi) \leftrightarrow (M \vdash \psi))$$

is true, and for any P there is such an M. Notice that this formulation of Mechanism requires us to restrict our attention to arithmetic sentences and what the human mind can (or cannot) prove in arithmetic.

1.3 A Key Assumption

The human mind (like a digital computer) is limited with respect to time, its memory, the amount of energy that it can allocate to proving arithmetic sentences, etc. The finitude of the human mind trivially guarantees that the set of arithmetic sentences which a person (P) can, *in fact*, prove, call it S, is finite. Given that S is finite, it would be a relatively easy task to construct a formal system that proves all and only the sentences in S and thereby exactly models P's mind. [3] Perhaps not surprisingly then, both Lucas's argument and Penrose's argument require one to consider not what the mind can, *in fact*, prove, but what the mind can prove *in principle* (it is less clear to me that McCall's argument needs this assumption [4]). That is, they each require S to be the set of arithmetic sentences provable by P if P's mind were not limited by time or memory or energy (etc.). I take it that considering the mind in these idealized terms has the effect (possibly among other things) that the

[3] One way of doing this would be to wait until P dies. Catalogue the arithmetic sentences that P proved during P's lifetime and then create a Turing machine which outputs those sentences one after another, and nothing else.

[4] One of the key points made by McCall in his argument is that the human mind is capable of knowing about the category of true but not provable sentences, but no formal system can know about that category (for my discussion of this see §4.1). It is not clear to me whether McCall *needs* to consider the mind in idealized terms to make this point.

idealized mind at least extends Peano Arithmetic (PA). [5] [6] It could be argued that the factual limitations on the mind make the notion of an 'idealized mind' described above somewhat contrived and even confusing. The notion is rather vague, which makes it unclear which formal system one would be exactly modeled by were one's mind idealized in the above way. [7] Hence, it may be difficult to see how this notion could be useful in assessing the plausibility of Mechanism. [8] Still, the Lucas-Penrose arguments require considering the mind in the idealized terms outlined above, so I will set such concerns aside and proceed with this notion in place. [9]

[5] By "the idealized mind at least extends PA" I mean that one can assume that an idealized subject is able to prove at least all of the statements provable in PA and possibly more.

[6] I think that it is worth noting that there is a slight difference between what it means to be an idealized subject (i.e. what an idealized subject can prove) within the context of Lucas's argument and what it means to be an idealized subject within the context of Penrose's argument. Within the context of Lucas's argument, S is the set of arithmetic sentences that are provable by a *particular* idealized mind. This is distinct from Penrose's account in which S should be understood as the set of arithmetic sentences that are provable by *any* human mind that is not limited by time or memory or energy, at any time in human history (past, present or future). I do not think that using Lucas's method for constructing S would have a too significant of an impact on Penrose's argument; however, Penrose's method for constructing S will not do for Lucas. As will be discussed in the next section, Lucas appeals to the dialectical nature of his argument and requires the mechanist to produce a formal proof of the consistency of the formal system which the mechanist alleges to be an exact model of Lucas's mind (see §2.4 for my discussion of this). Such a proof would be impossible (if Lucas's mind is consistent) if membership in S is defined as Penrose describes. Admittedly, some of Lucas's commentators have thought that he may define membership in S just as Penrose does (see, for example, (Shapiro 2003 pp. 25–26)), but I will opt for what I take to be the more charitable interpretation described above.

[7] Benacerraf gives an interesting and rather sensible way of understanding what an idealized mind is. He says S is the set of sentences that the non-idealized mind (under consideration) can prove and "S^* is the closure of S under the rules of first order logic with identity." (Benacerraf 1967 p. 24) The 'idealized mind' proves (every member of) S^*. I am not sure that this would be *idealized enough* for Lucas and Penrose, but I think that it is sensible.

[8] For a more detailed discussion of the problems associated with this notion of idealized minds, see (Franzén 2005 pp. 82–83); and (Shapiro 1998 pp. 275–277).

[9] There is another assumption made by Lucas and Penrose (and possibly McCall) which I do not think is directly relevant to *my* discussion of their arguments, but I will mention it

2 The Lucas Argument

In his paper, Minds, Machines and Gödel (1961), Lucas argues that Gödels incompleteness theorems show that Mechanism is false. Lucas contends that his ability to understand and apply Gödel's incompleteness theorems enable him to show that no formal system can prove all of the same arithmetical sentences that he can prove. Hence, no formal system can exactly model his mind. Therefore, Gödel's incompleteness theorems imply that Mechanism is false. In this section I will examine Lucass argument and show that it fails to demonstrate that Gödel's theorems are incompatible with Mechanism.

I have divided this section into four main subsections. In section (2.1), I present Lucas's Gödelean argument against Mechanism. In order for Lucas to show that Gödel's incompleteness theorems are incompatible with Mechanism, Lucas must (somehow) be able to prove the consistency of the formal system that is alleged to exactly model his mind (call this system, M), from which he could derive M's Gödel sentence. Moreover, Lucas must do this without

here. Explicit in Penrose, and I think implicit in Lucas (and possibly McCall), is the idea that correctable mistakes ought not to be considered in this discussion. Included among the 'correctable mistakes' are at least two types of mistakes. The first type of mistake is an error in the execution of some algorithm or procedure. A simple example of such a mistake would include forgetting to "carry the one" when adding 15 and 7. The second type of correctable mistake involves deploying questionable assumptions or rules when engaging in mathematical reasoning which may lead to paradox. As an example of this sort of thing, Penrose cites the paradox of "the set of all sets which are not members of themselves" which Bertrand Russell pointed out was a consequence of some of Gottlob Frege's work (Penrose 1996 p. 138). Correctable mistakes are to be omitted from this discussion or else, if they are allowed, the mind of the idealized subject is not to be impugned on the basis of such mistakes. That is, one is not permitted to conclude that the subject's mind is inconsistent because the subject makes correctable mistakes. This may raise the question, when is one permitted to conclude that the subject's mind is inconsistent? One may conclude that the idealized subject's mind is inconsistent just in case the subject's mind will not allow the subject to correct some contradiction. That is, the contradiction is fundamental or *built in*, as it were, to the subject's mind. No act of will would enable the subject to overcome the inconsistency.

compromising his own consistency. In section (2.2), I point out that Lucas does not actually (consistently) produce a formal (Lucas-)proof of the consistency of M and so the mere claim that Lucas is able to produce such a proof presents no obvious threat to Mechanism. In section (2.3), I argue that Lucas's ability to produce an *in*formal proof of the consistency M does not show that Gödels theorems are incompatible with Mechanism. In section (2.4), I argue that Lucas cannot make use of a formal proof (which has been produced in a non-Lucas-system) of the consistency of M in a way which demonstrates that Gödels theorems are incompatible with Mechanism.

2.1 Lucas's Argument

Lucas's argument may be put as follows:

(L1) Suppose Mechanism is true, for *reductio*.

(L2) For any consistent formal system, M, which a mechanist claims exactly models Lucas's M.

Let S_L be the set of arithmetical sentences that Lucas can prove and let S_M be the set of arithmetical sentences that M can prove.

(L3) M can formally prove the sentence, $Con_M \rightarrow G_M$ (where 'Con_M' states the consistency of M).

(L4) So, Lucas can formally prove $Con_M \rightarrow G_M$ by the assumption that M exactly models Lucas.

(L5) Since M is, in fact, consistent, Lucas can apply modus ponens to $Con_M \rightarrow G_M$.

(L6) So, Lucas can formally prove that G_M (is true).

(L7) So, G_M is in S_L.

(L8) M cannot formally prove that G_M (is true) by G1.

(L9) So, G_M is not in S_M

(L10) Hence, S_L and S_M are not identical sets by definition.

(L11) Therefore, M cannot be an exact model of Lucas's mind.

Lucas can run this argument for any M, so statement (L11) generalizes to: *no (particular) formal system can be an exact model of Lucas's mind.* Hence, Mechanism must be false.

It could be objected that Lucas's argument fails as a *reductio* against Mechanism because Lucas's mind, and the formal system that models it, could be inconsistent. If Lucas's mind is inconsistent, then the mechanist could specify an inconsistent formal system (call it, IM) as its exact model. Inconsistent systems are not subject to Gödels incompleteness theorems. Moreover, since every arithmetical sentence is derivable in an inconsistent system, the set of arithmetical sentences that Lucas can prove must be identical to the set of arithmetical sentences that IM can prove. Hence, Lucas's argument fails as a *reductio* against Mechanism because it fails to consider the possibility that Lucas's mind is inconsistent.

Lucas tries to avoid this objection, insisting that it is implausible to think that his mind (or the human mind generally) is inconsistent. Lucas appeals to empirical evidence in support of this position. He claims that humans do

not behave the same way that inconsistent formal systems operate. He writes, "[W]e eschew inconsistencies when we recognize themIf we really were inconsistent machines, we should remain content with our inconsistencieshappily [affirming] both halves of a contradiction. Moreover, we would be prepared to say absolutely anythingwhich we are not." (1961 p. 121) Lucas appears to be right about the fact that humans are not generally observed behaving in the same way that inconsistent formal systems operate. Humans are discriminating with respect to what they assert and they view contradictions as problematic. If the mind were an inconsistent formal system one would expect that humans would assert anything and be comfortable with contradicting themselves. Sane humans do not behave that way. Hence, it is implausible to think that the human mind is an inconsistent formal system, or that it could be exactly modeled by one.[10]

2.2 Lucas does not Formally Prove Con_M

I am sympathetic with Lucas's position; however, even if it is conceded that the empirical evidence suggests that it is implausible to deny that the human mind is consistent, [11] it is not clear that Lucas's argument succeeds at showing

[10](Lucas 1961 p. 121)

[11] I am not sure that the mechanist *should* concede this. There is a problem with Lucas's argument for the consistency of the mind that is worth mentioning. It is not clear that his argument should be admitted into the present discussion. One could charge that Lucas's evidence in support of the consistency of the mind cannot be used in light of the parameters set in place to make sense of his argument, to wit, considering the mind in idealized terms. It may be true that when considering idealized minds, the best explanation for apparently consistent behavior *is* consistency; however, it is not at all obvious that one actually *would* observe apparently consistent behavior were one actually observing an *idealized* mind. In the absence of a reason to suppose that one would observe apparently consistent behavior in an idealized subject, should one grant Lucas his *factual* evidence of apparently consistent behaviour of the mind, one may reason that the mind need no longer be considered in

that Gödel's theorems imply that Mechanism is false. As it stands, Lucas's argument is unsound because it contains an obviously false premise. Statement (L6) is the crucial premise in Lucas's argument and it is supposed to follow from statement (L5) (and (L4)). The trouble is that statement (L5) is clearly false. It is not enough that M is, in fact, consistent. In order for Lucas to soundly apply modus ponens to $Con_M \rightarrow G_M$ and thereby formally prove G_M, Lucas must first formally prove Con_M. Hence, in order for (L6) to follow, statement (L5) must be changed to (L5′) *Lucas can formally prove* Con_M. Statement (L6) does follow from (L5′) (and (L4)), but why should one think that (L5′) is true? Lucas does not actually give a formal proof of Con_M and (L5′) is not (sufficiently) justified by the mere *claim* that Lucas can formally prove Con_M. Given the absence of support for (L5′), the mechanist is justified in rejecting it.

The mechanist is justified in rejecting (L5′). Statement (L1) in Lucas's argument is the assumption that Mechanism is true. Statement (L2) stipulates that M is a consistent formal system which exactly models Lucas's mind. So from statements (L1) and (L2) the mechanist is entitled to assume:

(I) $\forall\psi((Lucas \vdash \psi) \rightarrow (M \vdash \psi)$

From (I) by substitution, the mechanist would know:

(II) $(Lucas \vdash Con_M) \rightarrow (M \vdash Con_M)$

The contrapositive of (II) is:

idealized terms. Accordingly, one could explain the apparently consistent behavior without appealing to the actual consistency of the mind. For example, it is possible that the proof of the inconsistency (of the mind) is so long that it is impossible to encounter it in a human life time. This could explain why one might *behave* as though one is consistent, even though one is, in fact, inconsistent.

(III) $(M \nvdash Con_M) \rightarrow (Lucas \nvdash Con_M)$

From G2 and the stipulation that M is consistent, the mechanist would know:

(IV) $(M \nvdash Con_M)$

From (III) and (IV) it follows that:

(V) $(Lucas \nvdash Con_M)$

This shows that if statements (L1) and (L2) in Lucas's argument are true, then it should be impossible for Lucas to formally prove Con_M. [12] This implies that (L5′) is at least as contentious as the conclusion that Lucas wishes to draw from it, namely the denial of Mechanism. Hence, unless Lucas gives some good reasons(s) in support of (L5′), adding (L5′) to Lucas's argument begs the question. Lucas does not give good reason(s) in support of (L5′). Therefore, adding (L5′) to Lucas's argument begs the question. Ergo, the mechanist is justified in rejecting (L5′).

There is another problem with (L5′) which is worth mentioning. Even if (L5′) were (demonstrably) true, the mechanist could still be justified in rejecting Lucas's argument as a refutation of Mechanism. Replace statement (III) above with:

(III*) $Lucas \vdash Con_M$ [13]

[12] Note: this argument does not show that it is outright impossible for Lucas to prove Con_M, but *only* that it is impossible *if* Mechanism is true (and M is consistent and an exact model of Lucas's mind). The truth of Mechanism is precisely what Lucas is arguing against and so asserting that it is outright impossible for Lucas to prove Con_M would amount to begging the question against *Lucas*. That is not my intent. My intent is only to show that in the absence of some good evidence in support of the claim that Lucas can formally prove Con_M, the mechanist is justified in rejecting that claim (assuming of course that the mechanist is justified in thinking that Mechanism is true).

[13] Statement (III*) amounts to statement (L5′).

From (II) and (III*) it follows that:

(IV*) $M \vdash Con_M$

(IV*) and G2 imply that M is inconsistent. Therefore, the mechanist can either reject the claim that M exactly models Lucas (thus avoiding (IV*)) and reject Mechanism, or conclude that Lucas's mind (and the formal system which exactly models it) must be inconsistent after all. Lucas does not provide sufficient reasons to prevent the mechanist from doing the latter. If the mechanist is justified in believing that Mechanism is true and the mechanist has a defeater for Lucas's reason(s) in support of the consistency of his mind (i.e. his empirical evidence), then, when confronted with a (Lucas-)proof of Con_M, the mechanist is justified in concluding that Lucas's mind must be inconsistent after all. The mechanist is justified in believing that Mechanism is true. [14]

Moreover, the mechanist has a defeater for Lucas's reason(s) in support of the consistency of his mind. The mechanist can argue that it is possible that the mind is, in fact, inconsistent but that (sane) humans generally *behave* as though they are consistent because the proof of the inconsistency (of the mind) is so long that it is impossible to encounter it in a human life time. [15] Hence, the mechanist is justified in believing that Mechanism is true and the mechanist has a defeater for Lucas's reason(s) in support of the consistency of his mind. Therefore, when confronted with a (Lucas-)proof of Con_M, the

[14] I take it as a given that the mechanist has reasons in support of Mechanism which justify the mechanist in believing that Mechanism is true.

[15] See footnote 12 for a more detailed explanation of why the mechanist could be justified in thinking that the mind is *inconsistent* despite the apparent empirical evidence to the contrary.

mechanist is justified in concluding that Lucas's mind must be inconsistent after all. Therefore, even if (L5′) were (demonstrably) true, the mechanist could still be justified in rejecting Lucas's argument as a refutation of Mechanism.

If Lucas's argument is to convincingly show that Gödel's theorems are incompatible with Mechanism, Lucas must explain not only how he is able to demonstrate Con_M, but also how he is able to carry out that demonstration while maintaining his own consistency. With respect to (L5′), Lucas does not do this (nor is it is obvious, to me, how this *could* be done).

2.3 Lucas and M can *In*formally Prove Con_M

If Lucas's argument is to succeed, Lucas must explain how he is able to demonstrate that Con_M (is true) and why that demonstration does not compromise his own consistency. One way Lucas might go about doing this is by claiming that statement (L6) in his argument is not supposed to follow from (L4) and (L5′), but from (L4) and (L5*) *Lucas can informally prove* Con_M. That is, Lucas can conclude that Con_M (is true) through empirical evidence or by some non-formalizable-in-M (or in-Lucas) means. Making this change seems to give Lucas the explanation that he needs. Lucas *has* provided an empirical argument which supports (L5*). [16] Moreover, the truth of (L5*) does not compromise Lucas's consistency because Lucas's producing an *in*formal proof of Con_M does not run afoul of G2 given the stipulation that M is, in fact, consistent (and under the assumption that M exactly models Lucas). [17]

[16] As I have mentioned, there is room for debate here. A mechanist may not find Lucas's empirical argument in favour of the consistency of the mind very compelling (see footnote 12).

[17] G2 precludes a consistent formal system from producing a *formal* proof of its own consistency. It says nothing of an *in*formal proof thereof.

The change from (L5′) to (L5*) would give Lucas the explanation that he needs, but, as Paul Benacerraf points out in, "God, The Devil, and Gödel" (1967), the change gives rise to new problems for Lucas's argument. Benacerraf observes that this change undermines the validity of Lucas's argument. The change makes Lucas's argument commit a sort of equivocation insofar as the sense of 'proves' must shift between '*in*formally proves' in statement (L5*) and 'formally proves' in the following premise (statement (L6)). This shift in the sense of 'proves' is made explicit in my formulation of Lucas's argument and so the problem with changing (L5′) to (L5*) is obvious. It does not follow from the fact that Lucas can *in*formally prove something that he can also formally prove it. Hence, statement (L6) does not follow from statement (L5*). Rather, what does seem to follow is statement (L6*), *Lucas can informally prove that* G_M *(is true).* Hence, in order to preserve the validity of his argument, Lucas should change statement (L6) to (L6*). Yet, if this change is implemented, then, as Benacerraf notes, "it is not at all clear that the prowess *claimed* for the mind is one that *Gödel I* [G1] precludes for machines." (1967 p. 19) His point is that Gödel's results do not imply that M cannot *in*formally prove Con_M, nor do they imply that M cannot *in*formally prove G_M. Hence, it is not obvious that Lucas's argument shows that he is able to outstrip M. To put it another way, if S_L (the set of arithmetical sentences that Lucas can prove) and S_M (the set of arithmetical sentences that M can prove) include the arithmetical sentences which Lucas and M can *in*formally prove, then Lucas's appeal to G1 (in statement (L8)) does not imply that G_M is not in S_M (i.e. statement (L9) in Lucass argument). Ergo, it does not follow that S_L and S_M are not identical sets. Therefore, Lucas cannot draw his conclusion

from (L5*).[18]

Lucas responds to Benacerraf's criticism by insisting that his ability to produce an informal proof of Con_M (and thereby of G_M) *is* sufficient to demonstrate, from G1 and G2, that he is not exactly modeled by M. Lucas argues that though *he* is able to carry out such a task, M could not "produce-as-true"[19] its own consistency, even *informally*, without violating G2. He notes that the operations of a Turing machine that was programmed to imitate informal proofs of the sort that Lucas uses to assert (or justify the assertion), Con_M, must "be governed by [its] programme, and would correspond to a formal system". (1968 p. 147) Lucas is saying that the operations of any Turing machine will correspond to the operations of some formal axiomatized theory. That is, for any Turing machine, whatever sentences it outputs as true, there is a formal axiomatized theory that formally proves all and only the exact same sentences. Hence, if M is a Turing machine that is able to "*in*formally prove" its own consistency, then the formal axiomatized theory that corresponds to M, [20] call it F_M, must also assert Con_M. Since F_M is a formal axiomatized theory, in order for it to assert Con_M, it must prove Con_M in the formal sense, but that violates G2 (if M is consistent) because Con_M and Con_{FM} (the consistency statement for F_M) would be logically equivalent. Ergo, M cannot, even informally, prove Con_M without violating G2 (or G_M without violating G1). [21]

There are two main problems with Lucas's contention that his ability to

[18] For Benacerraf's discussion of this see (Benecerraf 1967 pp. 19–20)

[19] (Lucas 1968 p. 147)

[20] That is, the formal axiomatized theory that formally proves all and only the same arithmetic sentences that M outputs.

[21] For this argument in its entirety see (Lucas 1968 pp. 147–148).

informally prove Con_M is sufficient to draw the conclusion that he cannot be exactly modeled by M. Firstly, it is not convincing. Lucas claims that no formal system has the ability to produce an informal proof of its own consistency (without violating G2). Lucass procedure for informally proving Con_M is roughly as follows:

(i) Lucas can informally prove the consistency of his own mind (on empirical grounds).

(ii) By hypothesis, M exactly models Lucas.

(iii) So, if Lucas is consistent, then M is consistent.

(iv) Therefore, Lucas can informally prove the consistency of M.

Step (i) is problematic. Step (i) can be interpreted as indicating that Lucas can produce a formal proof of his own consistency [22] or that he can produce an *in*formal proof of his own consistency. Suppose step (i) indicates that Lucas can produce a formal proof of his own consistency. If Lucas's argument is to convincingly show that Mechanism is false, then Lucas must explain how he is able to execute step (i) without compromising his own consistency. Otherwise, when presented with a formal (Lucas-)proof of Lucas's consistency the mechanist is justified in concluding that Lucas's mind is inconsistent. [23] Lucas does not explain how he is able to execute step (i) without compromising

[22] Admittedly, interpreting step (i) as indicating that Lucas can produce a formal proof of his own consistency seems like quite a stretch. However, I think that Lucas's claim that an informal proof carried out in formal systems is still a kind of formal proof gives the mechanist license to explore the above as a possible interpretation of step (i).

[23] The reason for this is analogous to the reason that the mechanist is justified in concluding that Lucas's mind is inconsistent when presented with a (Lucas-)proof of Con_M (see §2.2).

his own consistency. Therefore, Lucas's argument does not convincingly show that Mechanism is false (under the supposition that step (i) indicates that Lucas can produce a formal proof of his own consistency). Alternatively, suppose step (i) indicates that Lucas can produce an *in*formal proof of his own consistency. It follows that Lucas must possess an ability which he claims no formal system can possess, namely, the ability to informally prove his own consistency. If Lucas has this ability, then it is the fact that he possesses this ability which shows that mechanism is false and not strictly his ability to prove Con_M and G_M (which is what is required for Lucas's argument to show that Gödels theorems are incompatible with Mechanism). Worse still, affirming that Lucas has the ability to informally prove his own consistency (and that no formal system has that ability) could be construed as begging the question against Mechanism.

The second problem with Lucas's contention that his ability to informally prove Con_M is sufficient to draw the conclusion that he cannot be exactly modeled by M is that it is false. Lucas supports his contention by (essentially) arguing M cannot informally prove Con_M without violating G2 because there is no real difference between an *in*formal proof and a *formal* proof in a formal system. However, this is inaccurate. It seems perfectly reasonable that a formal system could produce a statement as true with less than absolute certainty attached (i.e. *probably* true). There are, for example, Turing machines which calculate probabilities. Perhaps, Lucas's mind is a formal system which is able to calculate probabilities. It takes as input observations of human behavior, combines this with knowledge of the way in which inconsistent formal systems operate and calculates the probability that the human mind is inconsistent

to be low. From this, Lucas could conclude with a high degree of certainty (but less than absolute certainty) that his mind is consistent. [24] Lucas thus demonstrates that 'it is highly probable that-Con_M' and so, by modus ponens on $Con_M \rightarrow G_M$, no more than, 'it is highly probable that-G_M', but M could also carry out this procedure without violating G1 or G2. Hence, there is a difference between an informal and a formal proof in a formal system and, moreover, it *is* possible for M to informally prove both Con_M and G_M without violating G1 or G2. Therefore, Lucas's ability to informally prove Con_M (and G_M) does not show that Lucas is able to "produce-as-true" a statement which M cannot also "produce-as-true". [25] Therefore, adding statement (L5*) to Lucas's argument and changing statement (L6) to (L6*) renders Lucas's argument ineffective as a demonstration that Gödel's theorems are incompatible with Mechanism.

[24] Notice that Lucas has formally proved, from the information available, that 'it is highly probable that his mind is consistent', but he has not formally proved that 'his mind is consistent'. This difference is important.

[25] Lucas may insist that when he asserts that his mind is consistent he is not claiming merely that it is highly probable that his mind is consistent, but that it is certainly consistent. His observations of human behavior combined with his knowledge of the way in which inconsistent systems operate convince him, with absolute certainty, that the human mind cannot be inconsistent. Indeed, this does *seem* to be what Lucas thinks, though this view is peculiar in light of his statements about simple arithmetic. He writes, "[I]n spite of our great feeling of certitude that our system of whole numbers which can be added and multiplied together is never going to prove inconsistent. It is just as conceivable we might find we had formalized it incorrectly." (1967 p. 123) Lucas thinks that it *is* possible that simple arithmetic may turn out to be inconsistent, despite the fact that it appears to operate consistently and despite the degree to which mathematicians may be certain that it is consistent. Why then would he insist, on the basis of, arguably less sturdy ground, that one can know with absolute certainty that one's mind is consistent? He should not. Indeed, as George Boolos points out, those two assertions appear irreconcilable (Boolos 1968 p. 614).

2.4 Lucas Cannot Use a Mechanist-Proof of Con_M (in a way that M Cannot)

Lucas thinks that his argument still succeeds as a refutation of Mechanism despite the criticisms presented in §§2.2–2.3. Lucas argues that the dialectical nature of his argument is integral to its success. He reasons that since the mechanist is claiming that M exactly models his mind, the mechanist must be familiar enough with M to know whether M is consistent or inconsistent. Thus, it is reasonable for Lucas to ask the mechanist if M is consistent. If the mechanist says, "*no*," then Lucas may reject the mechanist's claim as implausible. [26] Alternatively, if the mechanist says, "*yes*," then this gives Lucas the premise that he needs to apply modus ponens to $Con_M \rightarrow G_M$, thereby enabling him to formally prove G_M. [27] [28]

Lucas thinks that the onus to prove Con_M is on the mechanist. He is not explicit about whether he expects the mechanist to formally (mechanist-)prove Con_M or to *in*formally (mechanist-)prove Con_M, though he likely expects the former. If Lucas were expecting the mechanist to *in*formally prove Con_M, then his argument would have trouble getting off the ground. If the mechanist gives Lucas an informal proof of Con_M, then when Lucas applies modus ponens to

[26] Recall that Lucas thinks that given certain empirical evidence it is implausible to think that the mind is inconsistent (see §2.1).

[27] Lucas 1968 p. 154)

[28] I should note that it is rather odd that Lucas thinks that the mechanist's testimony can be made part of a formal (Lucas-)proof of anything. As Shapiro notes, "[T]his "word of the mechanist" does not give Lucas a premise he can use (in a modus ponens on $[Con_M \rightarrow G_M]$) simply because this word does not amount to mathematical certainty." (2003 p. 25) I think that Shapiro is ultimately right about this. Invoking someone's testimony in a formal proof is generally impermissible. At best, it seems that Lucas would be able to use the mechanist's testimony to produce an *in*formal proof of Con_M. However, as was discussed in §2.3, an *in*formal proof of Con_M is insufficient to show that Gödel's theorems imply that Mechanism is false. Still, I will give Lucas the benefit of the doubt and explore this possibility.

$Con_M \rightarrow G_M$, he will at most *in*formally prove G_M. To give a formal proof of G_M Lucas needs a formal proof of Con_M. Since Lucas needs to give a formal proof of G_M in order to outstrip M, Lucas must be expecting the mechanist to give him a formal proof of Con_M.

Initially, one may object to Lucas's claim that the onus to formally prove Con_M is on the mechanist on the grounds that it seems unwarranted. Why should the mechanist be expected to produce a formal proof of Con_M for Lucas? Lucas reasons that since the mechanist claims that M exactly models his mind, the mechanist must be familiar enough with M to know (i.e. prove) whether M is consistent or inconsistent. But why think that? The mechanist may be familiar enough with M to *know* that M is an exact model of Lucas's mind but M may be so complex that the mechanist has been unable to produce a formal proof of Con_M. Still, the mechanist may have strong *in*formal reasons for thinking that Con_M is true. Alternatively, perhaps the mechanist is unable to formally prove Con_M because M is an exact model of the mechanist's mind too. [29] It would be quite unreasonable of Lucas to expect the mechanist to do the impossible. [30] It is conceivable that the mechanist is familiar enough with M to know that M exactly models Lucas's mind, but that the mechanist is unable to formally prove Con_M. Hence, Lucas's claim that the onus to formally prove Con_M is on the mechanist seems unwarranted.

Lucas could potentially avoid the objection that his claim that the onus to formally prove Con_M is on the mechanist seems unwarranted by imposing some assumptions. Assume that the mechanist is not exactly modeled by M

[29] If Mechanism is true, and M is an exact model of the mechanist's mind, then it would be impossible for the mechanist to prove Con_M (if M is consistent).

[30] Impossible, that is, if Mechanism is true.

and that the mechanist is very clever and knows that M is consistent and can formally prove Con_M. These assumptions eliminate potential reasons why the mechanist may be unable to formally prove Con_M. Indeed, under these assumptions, the mechanist can formally prove Con_M (i.e. there is a formal (mechanist-)proof of Con_M) and so Lucas's request for the mechanist to produce the proof (of Con_M) is reasonable (and warranted).

I will grant that Lucas's request for the mechanist to produce a formal proof of Con_M is reasonable (and warranted) and so I will proceed under the assumptions that the mechanist is not exactly modeled by M and that the mechanist is clever and knows that M is consistent and can formally prove Con_M. These assumptions raise at least two related problems with Lucas's contention that once the mechanist gives him a formal proof of Con_M, he can soundly apply modus ponens to $Con_M \rightarrow G_M$ and thereby falsify Mechanism. I will consider each below.

2.4.1 Problem 1

Under the current assumptions, it is not obvious that Lucas could use the mechanist's proof of Con_M (or 'Con_M' once mechanist-proved) in a formal (Lucas-)proof of G_M. To explain why first requires briefly considering what it means to be a 'formal Lucas-proof' and what it means to be a 'formal mechanist-proof'. The Lucas-proof relation, 'is a Lucas-proof of '(call this relation, Prf_{Luc}) is constructed as follows. A sequence of formulae, α, is a formal Lucas-proof of some sentence, λ, just in case every formula in α is either an axiom of Lucas's system (mind) or follows from a previous formula by (at least) one of the inference rules of Lucas's system and λ is the last

formula of α. Hence, it is true that Prf_{Luc} holds between $\ulcorner\alpha\urcorner$ and $\ulcorner\lambda\urcorner$ (i.e. $Prf_{Luc}(\ulcorner\alpha\urcorner, \ulcorner\lambda\urcorner)$ is true) just in case α is a formal Lucas-proof of λ. The mechanist-proof relation, 'is a mechanist-proof of ' (call this relation, Prf_{Mec}) is constructed in a likewise manner only by the axioms and inference rules of the *mechanist's* system (mind).

The current assumptions imply that Lucas must hold that Prf_{Luc} and Prf_{Mec} are different relations. By assumption, M is, in fact, consistent and there is a formal mechanist-proof of Con_M. By G2, M cannot prove Con_M and so the axioms or inference rules of the mechanist's system must be distinct from the axioms or inference rules of M. From the hypothesis that M exactly models Lucas (i.e. statement (L2)) it follows that the axioms or inference rules of the mechanist's system must be distinct from the axioms or inference rules of Lucas's system . Of course, Lucas's argument is supposed to show that it is *false* that he is exactly modeled by M, so he may deny that that hypothesis can be used to show that his axioms or inference rules are distinct from the mechanist's. This may be true, but if Lucas wishes his argument to be persuasive, the assumptions in play still require Lucas to hold that Prf_{Luc} and Prf_{Mec} are different relations. Holding that Prf_{Luc} and Prf_{Mec} are the *same* relation begs the question. The assumptions that there is a mechanist-proof of Con_M and that M is in fact consistent, imply that the mechanist is not exactly modeled by M. Hence, holding that Prf_{Luc} and Prf_{Mec} are the same relation, presupposes that Lucas is not exactly modeled by M. Therefore, if Lucas wishes his argument to be persuasive, he must hold that Prf_{Luc} and Prf_{Mec} are different relations.

Given that Lucas must hold that Prf_{Luc} and Prf_{Mec} are not the same re-

lation, it is not clear why Lucas thinks that being given the mechanist's proof of Con_M could enable him to produce a formal Lucas-proof of G_M. When the mechanist proves Con_M for Lucas, the mechanist demonstrates to Lucas that the sequence of formulae α, is a formal mechanist-proof of Con_M. That is, the mechanist demonstrates to Lucas that $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M \urcorner)$ is true. In order for Lucas's argument to succeed, Lucas must be able to use this, in some way, to construct a formal Lucas-proof of G_M. Lucas does not explain how he intends to use $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M \urcorner)$ in his own formal proof of G_M. Presumably, Lucas still intends to formally prove G_M by modus ponens on $Con_M \rightarrow G_M$. This means that Lucas must be using $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M \urcorner)$ to derive Con_M in his system. But how can he do that? Lucas cannot use α to formally prove Con_M for himself (i.e. in his system). Prf_{Luc} and Prf_{Mec} are different relations and so, it is invalid for Lucas to infer $Prf_{Luc}(\ulcorner\alpha\urcorner, \ulcorner Con_M \urcorner)$ from $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M \urcorner)$. Perhaps Lucas plans to infer Con_M directly from $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M \urcorner)$. This would be odd. It requires postulating that Lucas's system contains a rather dubious inference rule something like: *if φ is a theorem of the mechanists system, infer φ.* Not only is this rule strange and *ad hoc*, it also begs the question. The rule enables Lucas to prove all of the arithmetical sentences which the mechanist can prove. By assumption the mechanist is not exactly modeled by M. Therefore, Lucas is not exactly modeled by M. Hence, postulating that Lucas's system has this strange inference rule presupposes that Lucas is not exactly modeled by M. Going this route would beg the question (it would also make Lucas's proof of G_M superfluous). [31] Therefore, Lucas must not intend to infer Con_M directly from

[31] It is perhaps worth noting that Lucas could avoid the problem of question begging here

$Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M\urcorner)$. Failing the above two options, it is not at all clear how Lucas could use $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M\urcorner)$ to produce a Lucas-proof of G_M.

2.4.2 Problem 2

Lucas could respond to Problem 1 by arguing that it is not his knowledge *that* $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M\urcorner)$ is true which enables him to produce a Lucas-proof of Con_M, rather it is his knowledge of *why* $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M\urcorner)$ is true. That is, once Lucas sees the mechanist's proof of Con_M and examines how it was carried out (i.e. which axioms and inference rules are needed for α to be a valid formal proof of Con_M), Lucas can modify his system (mind) such that α becomes a Lucas-proof of Con_M. Lucas simply recognizes which axioms and inference rules are required to make $Prf_{Mec}(\ulcorner\alpha\urcorner, \ulcorner Con_M\urcorner)$ true and then adds the required axioms or rules of inference to his system such that α becomes a Lucas-proof of Con_M. Once Lucas has done this he will be able to apply modus ponens to $Con_M \rightarrow G_M$ and thereby generate a formal Lucas-proof of G_M. Thus, Lucas can prove that he is not exactly modeled by M and thereby show that Gödel's theorems are incompatible with Mechanism.

There is a difficulty with the above response to Problem 1. If Lucas modifies his system in order to formally (Lucas-)prove Con_M, then the terms of the discussion change such that Lucas's argument does not show that Gödel's results imply that Mechanism is false. In "Lucas Against Mechanism II" (1979), David Lewis makes this point by arguing along the following lines. A *modi-*

by dropping the assumption that the mechanist is not exactly modeled by M. However, this would raise other problems for Lucas's argument. Particularly, it would prevent Lucas from being able to avoid the objection that his request that the mechanist provide him with a formal proof of Con_M is unreasonable (and unwarranted).

fied Lucas, call him, Lucas*, is not identical to the unmodified Lucas, Lucas. Hence, Lucas* is not exactly modeled by M and so Lucas* can prove Con_M and thereby derive G_M without contradiction. M will either possess the ability to modify itself in the same way that Lucas can modify himself, or it will not. If M does not have this ability, then it is Lucas's ability to modify his system that distinguishes him from M and not strictly his ability to prove G_M. [32] If M *can* modify itself, as Lucas can, then M, if presented with the mechanist's proof of Con_M, will modify itself to M^* an exact model of Lucas*. M^* will be different from M and so no contradiction will ensue if M^* can prove Con_M. Furthermore, since M^*, like Lucas*, can prove $Con_M \rightarrow G_M$, M^* can *also* prove G_M without violating Gödel's theorems. Consequently, Lucas would now need to show that S^*_L (the set of arithmetical sentences provable by Lucas*) is not identical to S^*_M (the set of arithmetical sentences provable by M^*) in order to falsify Mechanism; however, his argument fails to establish that that is the case. Therefore, if Lucas modifies his system in order to formally Lucas-prove Con_M, then his argument does not show that Gödel's results imply that Mechanism is false. [33]

There does not appear to be a way for Lucas to adapt the mechanists formal proof of Con_M into a formal Lucas-proof of Con_M (from which Lucas could derive G_M) in such a way as to show that Gödel's theorems are incompatible with Mechanism. Lucas's claim that the mechanist's formal proof of Con_M gives him the premise he needs to apply modus ponens to $Con_M \rightarrow G_M$ is either false, or requires Lucas to modify his system. In either case, Lucas's

[32] Claiming that Lucas has the ability to modify his system but denying that M has the same ability could be construed as begging the question against Mechanism.

[33] For Lewis's version of this argument see (Lewise 1979 p. 376)

argument fails to show that Gödel's theorems are incompatible with Mechanism.

2.5 Concluding Remarks on Lucas's Argument

Lucas's argument fails to demonstrate that Gödel's incompleteness theorems are incompatible with Mechanism. In order for Lucas's argument to run, Lucas must somehow (consistently) prove Con_M, from which he could derive G_M and thereby demonstrate that he is not exactly modeled by M. I examined three ways Lucas may attempt to execute this task. First, Lucas could formally (Lucas-)prove Con_M. As was discussed, Lucas does not provide a formal (Lucas-)proof of Con_M and the mere *claim* that Lucas can formally (Lucas-)prove Con_M can be rejected as it begs the question. Moreover, Lucas does not give a compelling reason to suppose that he even could produce a formal (Lucas-)proof of Con_M without compromising his own consistency. Hence, if Lucas *were* to produce a formal (Lucas-)proof of Con_M, the mechanist is entitled to conclude that Lucas's mind is inconsistent. Second, Lucas may attempt to prove Con_M by means of an *in*formal proof. It was shown that, though Lucas could consistently produce an *in*formal proof of Con_M, M could do the same thing (without violating G1 or G2). Hence, Lucas's ability to informally prove Con_M does not show that he is not exactly modeled by M or that Gödel's theorems are incompatible with Mechanism. Third, Lucas could attempt to use a formal (mechanist-)proof of Con_M in his own formal proof of G_M. It was shown that this seems possible only if Lucas modifies his system such that it contains the axioms or inference rules used by the mechanist to produce the formal (mechanist-)proof of Con_M (thus making the mecha-

nist's proof into a formal (Lucas-)proof). Yet, M could also carry out this procedure without violating G1 or G2. Hence, Lucas's ability to use a formal (mechanist-)proof of Con_M in a formal (Lucas-)proof of G_M does not show that Gödel's theorems are incompatible with Mechanism. Therefore, Lucas's argument fails to show that Gödel's incompleteness theorems are incompatible with Mechanism.

There are serious difficulties with Lucas's attempt to use Gödel's theorems to show that he cannot be exactly modeled by any formal system. It would seem, therefore, that if one wishes to construct a convincing argument to show that Gödel's theorems are incompatible with Mechanism, one should try a different approach. Roger Penrose does precisely that. In the next section I will examine Penroses attempt to demonstrate that Gödel's incompleteness theorems can be used to show that the assumption that he is exactly modeled by some formal system leads to contradiction.

3 The Penrose Argument

In *Shadows of the Mind* (1996), Roger Penrose devotes a considerable amount of space to the development of Gödelean arguments against Mechanism. Below I will examine the second Gödelean argument that Penrose presents in *Shadows of the Mind* (often dubbed in the literature as "Penrose's new argument"). Penrose attempts to overcome some of the more problematic elements in the Lucas argument by structuring his argument in a slightly different way. Penrose argues that he will be able to deduce certain sentences from the assumption that he is exactly modeled by some formal system, M. Penrose

argues that the assumption that he is exactly modeled by M together with Gödel's incompleteness theorems (and the fact that he can know that his mind is sound) allow him to deduce a contradiction. He concludes that Gödel's theorems imply that Mechanism is false. In this section I will show that Penrose's argument fails to establish that Gödel's incompleteness theorems are incompatible with Mechanism.

I have divided this section into four main subsections. In section (3.1), I present Penroses argument. Penroses argument relies, crucially, on Penrose's ability to know that his mind is sound. In section (3.2), I present a sketch of Penrose's argument for his (knowledge of) his own soundness and a worry for that argument. In section (3.3), I develop Penrose's argument for his own soundness in a bit more detail and present some criticisms. In section (3.4), I argue that even if Penrose were able to know that his mind is sound his argument would fail to run.

3.1 Penrose's Argument

To show that Gödel's theorems are incompatible with Mechanism, Penrose reasons as follows:

(P1) Suppose I am exactly modeled by a formal system, M (or "My mind is exactly modeled by M" is true).

(P2) I know that my mind is sound

(P3) So, I know that M is sound. (from P1, P2)

Let, M^*, be the formal system M plus the statement: "My mind is exactly modeled by M". [34] That is, M^* is the formal system which outputs a set of sentences, S_{M^*}, such that S_{M^*} contains every sentence output by M plus the sentence "My mind is exactly modeled by M" and no other sentences.

(P4) I know M^* is consistent. (because M is sound and "My mind is exactly modeled by M" is true by assumption. So S_{M^*} is sound. Ergo, M^* is consistent.)

(P5) I am exactly modeled by M^*. (because, by assumption, I am exactly modeled by M and I know "My mind is exactly modeled by M")

(P6) So, M^* knows that M^* is consistent. (from P4, P5)

(P7) M^* cannot know that M^* is consistent (by G2)

(P8) Contradiction.

Since statement (P1) leads to contradiction, Penrose concludes that it must be false. Penrose cannot be exactly modeled by M. Moreover, since M is arbitrary, this conclusion generalizes to: *no formal system can exactly model Penrose's mind*. Therefore, Mechanism must be false. [35]

3.2 Penrose's Argument for his Soundness

Initially, it seems that statement (P2) is the most controversial statement in Penroses argument. In order to draw the contradiction from Gödel's theorems

[34] Penrose does not specify which language the statement "My mind is exactly modeled by M" is supposed to be expressed in.

[35] Penrose gives summaries of this argument in (1996 p. 187) ; and (1996-B §3.2).

that he needs, the sense of the word 'knows' which Penrose has in mind needs to be something like 'formally proves'. Yet, if this is the intended sense of 'knows', then it is difficult to see how Penrose could *know* that his mind is sound without contradicting himself or else begging the question against Mechanism.

Penrose thinks that statement (P2) is reasonable, and demonstrable, given the parameters of his argument. Penrose's argument deals with a system (in this case Penrose's mind) reasoning about its own beliefs, specifically, what he calls "unassailable" beliefs. Penrose insists that there is a certain class of statements which are "unassailable" and that it is implausible to think that such statements are false. If Penrose only asserts statements which are "unassailable" it follows that it is implausible that his mind is unsound. But when does a statement count as "unassailable"? Penrose is not that specific, he writes:

> Perhaps mathematicians have now become more cautious as to what they are prepared to regard as 'unassailably true'—after a period of excessive boldness...at the end of the nineteenth century...I am happy that only those things whose truth is indeed unassailable should be included in the discussion, and that anything concerning infinite sets that is at all questionable should not be so included. The essential point is that *wherever* the line is drawn, Gödel's argument produces statements that remain within the compass of what is indeed unassailable...Doubtful issues in relation to the kind of very free reasoning that Cantor, Frege, and Russell were concerned with need not concern us so long as they remain 'doubtful' as op-

> posed to 'unassailable'. This being accepted, I cannot really see that it is plausible that mathematicians are *really* using an *unsound* formal system F as the basis of their mathematical understandings and beliefs. (1996 pp. 139–140)

Penrose seems to be saying that a statement is "unassailable" just when it is believed to be true with an extremely high level of conviction or certainty. Perhaps something on the order of the level of conviction that a mathematician might have in believing some theorem to be true after the mathematician has proved it. The key is that "unassailable" statements are beyond doubt and denying their truth is implausible. Penrose limits the beliefs that he can assert to only statements within the class of statements which are (believed to be) "unassailable". Hence, there is no reason to doubt the soundness of Penrose's mind. Therefore, Penrose argues, he knows that his mind is sound and statement (P2) is true.

There is at least one worry with Penrose's appeal to the "unassailable" in his argument in support of statement (P2). Even if Penrose limits all of the sentences which he asserts to the "unassailable" it is not obvious that he could use that fact in a formal proof of his own soundness. There are two reasons for this. First, *unassailability* seems like an epistemic property or relation. Penrose describes sentences as being "unassailable" when they are *not doubtful* and when they are such that it is implausible to deny their truth. This suggests that *unassailability* is, in a sense, a measure of Penrose's level of certainty or the strength of his epistemic position with respect to certain sentences. It does not necessarily follow from the fact that Penrose is certain that a sentence is true that that sentence *is*, in fact, true. Second, *unassailability* seems too

subjective to be used in a formal proof of soundness. Penrose thinks that the criteria which determine whether a sentence is "unassailable" can vary from system to system. [36] He writes,

> If our robot is really to have capabilities, understandings, and insights of a human mathematician, it will require some kind of concept of 'unassailable mathematical truth'...If our robot is to behave like a genuine mathematician, although it will still make mistakes from time to time, these mistakes will be correctable—and correctable, in principle, according to its *own* internal criteria of 'unassailable truth'. (1996 pp. 157–158)

This indicates that Penrose does not think that the criteria for *unassailability* need to be the same for every system. Since the criteria for *unassailability* are allowed to vary from system to system, it is conceivable that a sentence may be deemed "unassailable" by one system while another system deems that sentence's denial to be "unassailable". This cannot be done with *truth*. Hence, it is not obvious that *unassailability* can be used as a reliable indicator of truth. If Penrose is to use *unassailability* in a formal proof of his own soundness, then it needs to be a reliable indicator of truth. Therefore, it is not obvious that Penrose can use *unassailability* in a formal proof of his soundness.

Penrose does not offer a response to the worry that it is not obvious that he can use *unassailability* in a formal proof of his own soundness. This could be because he is unaware that it is a worry, or perhaps because he thinks that the structure of his argument eliminates it somehow. [37] Whatever the case

[36] By 'system' here, I mean either a formal system or a mind.

[37] I am not entirely sure how the structure of Penrose's argument *could* eliminate the

may be, neither Penrose nor his commentators make much ado about whether *unassailability* is the sort of thing that Penrose can use in a proof of his own soundness. Though I think it is a rather serious worry, I will set it aside and proceed with an examination of Penrose's argument *as though* it is not unclear that Penrose can use *unassailability* as a reliable indicator of truth. I suspect this is what many of Penrose's commentators are doing also.

3.3 Can Penrose Prove that He is Sound?

Penrose claims that he can prove the soundness of his own mind (i.e. statement (P2) is true) because he is able to demonstrate that every arithmetic statement he proves is "unassailable". If Penrose is using the property in a proof, then there must be a (1-place) predicate $U(x)$, in whatever language Penrose uses to produce proofs (call that language, L_P) such that an L_P-sentence φ, is "unassailable" just in case $U(\ulcorner\varphi\urcorner)$ is true. For Penrose to know that a sentences being "unassailable" guarantees that that sentence is not false, Penrose must know that:

(i) $U(\ulcorner\varphi\urcorner) \rightarrow \varphi$

is true for all L_P-sentences, φ. In order to use this to demonstrate his own soundness, Penrose must also know that he does not assert (that is, *prove*) any sentence if that sentence is not "unassailable". Or equivalently, Penrose

worry. Again, Penrose's argument deals with Penrose reasoning about his own beliefs. Perhaps Penrose thinks that this makes his certainty of the truth of the sentences which he proves particularly relevant to whether or not he believes himself to be sound. Or, perhaps Penrose thinks that restricting the context to only him reasoning about his own beliefs would avoid worries associated with the subjectivity of *unassailability*. As long as *he* does not assert that both a sentence and its denial are "unassailable" he can use the notion to guarantee his own soundness. This is not very convincing.

must know:

(ii) $\forall\varphi(Prov_P(\ulcorner\varphi\urcorner) \rightarrow U(\ulcorner\varphi\urcorner)$

where '$Prov_P$' is the standard provability predicate for Penrose. [38] In other words, Penrose is claiming that he has knowledge that (i), *for all L_P-sentences, if a sentence is "unassailable" then it is true* and that (ii) *all of the L_P-sentences which Penrose proves (or for which there is a Penrose-proof) are "unassailable"*. From his knowledge that (i) and (ii) it follows that Penrose knows that his mind is sound. Therefore, statement (P2) is true.

As has been pointed out by David Chalmers and others, the main problem with justifying statement (P2) in the above way, is that it leads to a contradiction which forces one to conclude either that statement (P2) is false, or else that Penrose's mind is inconsistent. [39] Whatever else one may wish to assume about the mind of an idealized Penrose one may be confident in assuming that it must be at least powerful enough to prove the diagonalization lemma. This means that there is an L_P-sentence, λ, such that Penrose proves:

(*) $\neg U(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$

Since Penrose knows that (i) is true for all L_P-sentences and λ is an L_P-sentence, λ can be substituted for φ in (i), and so Penrose must also prove:

(**) $U(\ulcorner\lambda\urcorner) \rightarrow \lambda$

From (*) and (**) it follows that $(U(\ulcorner\lambda\urcorner) \vee \neg U(\ulcorner\lambda\urcorner)) \rightarrow \lambda$. Since $U(\ulcorner\lambda\urcorner) \vee \neg U(\ulcorner\lambda\urcorner)$ is a tautology, Penrose proves it. Hence, Penrose must prove λ.

[38] $Prov_P(\ulcorner\varphi\urcorner)$ expresses $\exists x(Prf_{Penrose}(x, \ulcorner\varphi\urcorner))$

[39] See: (Chalmers 1995 §§3.8–3.13); (Shapiro 2003 p. 30); and (Lindström 2006 p. 235)

If Penrose proves λ, then there is a Penrose-proof of λ, and so from (ii) it follows that Penrose proves $U(\ulcorner\lambda\urcorner)$. Yet, from the fact that Penrose proves (*) and λ, it also follows that Penrose proves $\neg U(\ulcorner\lambda\urcorner)$. Hence Penrose proves a contradiction. Penrose either knows that his mind is sound or he does not. If Penrose knows that his mind is sound (through knowing (i) and (ii)), then it would seem that his mind is, in fact, inconsistent. Of course, if Penrose's mind is inconsistent, then his argument fails to show that Gödel's theorems are incompatible with Mechanism. [40] Alternatively, if Penrose does not know that his mind is sound, then statement (P2) is false and his argument against Mechanism is unsound.

Penrose responds to the criticism that he cannot know his own soundness because (i) and (ii) lead to contradiction by imposing a restriction and making a change. He writes:

> The belief system B, in question, is simply the one which "believes" (and is prepared to assert as "unassailably perceived") a [Π_1-sentence] S if and only if S happens to be true. B is not allowed to "output" anything other than a decision as to whether or not a suggested [Π_1-sentence] is true or false...However, as part of its internal musings, it is allowed to contemplate other kinds of thing[s].[41] (1996-B §3.8)

Penrose intends to impose a restriction and make a slight change. Firstly, each member φ of the set of arithmetic sentences which Penrose asserts (or proves),

[40] This is because his argument would fail to show that he can know something that no formal system can know. If Penrose's mind is inconsistent then so are M and M^*. Hence they each prove every arithmetic statement.

[41] Penrose uses the term "P-sentence" to refer to Π_1-sentences in the original text. I have changed 'P-sentence' to 'Π_1-sentence' throughout the quotation.

S_P, must be Π_1. [42] Secondly, change the claim that Penrose knows that (i) is true for all L_P-sentences to: *Penrose knows that*:

(i*) $U(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$

is true for all L_P-sentences φ if φ is Π_1. By imposing this restriction and making this change, the objection presented above is avoided and Penrose is still able to know that he is sound. Given the change, the L_P-sentence λ, for which it is provable that $\neg U(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$, will not be Π_1. This means that λ cannot be validly substituted into (i*). Moreover, since membership in S_P is restricted to only Π_1-sentences, Penrose cannot prove λ. Hence, no contradiction follows from (i*) and (ii). Restricting membership in S_P to only Π_1-sentences means that, if Penrose can prove that (i*) is true for all L_P-sentences φ if φ is Π_1 and that (ii) is true, then he can prove that S_P is sound. Penrose asserts that he can prove that (i*) is true for all L_P-sentences φ if φ is Π_1 and that (ii) is true. If so, Penrose knows that his mind is sound and statement (P2) is true.

Penrose's method for demonstrating his own soundness is no longer obviously contradictory, but it is hardly convincing. Penrose does not actually demonstrate that that he can prove that (i*) is true for all L_P-sentences φ if φ is Π_1 and that (ii) is true, he merely stipulates it. Given the restriction that Penrose can only prove Π_1-sentences, if (i*) is true for all L_P-sentences φ if φ is Π_1 and (ii) is true, then Penrose can only prove true (Π_1) sentences. Hence, Penrose is, in effect, "demonstrating" that he is sound by stipulating that he is only able to prove true (Π_1) sentences. *Of course* Penrose is sound if he is

[42] A Π_1-sentence is (or is logically equivalent to) a sentence which begins with one or more unbounded universal quantifiers (followed by a Δ_0 wff). It is important to note that 'Con_M' is a Π_1-sentence. 'Con_M' expresses '$\forall x \neg(Prf_M(x, \ulcorner\bot\urcorner))$'

only able to prove true (Π_1) sentences, but why should the mechanist grant that Penrose *is* only able to prove true (Π_1) sentences? Penrose's merely stipulating that he is only able to prove true (Π_1) sentences and claiming to know that he is sound on that basis is hardly convincing. Hence, the mechanist does not need to concede that (P2) is true. But things are even worse for Penrose because even *with* (P2) his argument fails.

3.4 Knowing He is Sound Does Not Save Penrose's Argument

Grant, for the moment, that Penrose knows that (i*) is true for all L_P-sentences φ if φ is Π_1 and that (ii) is true. This combined with Penrose's restriction that S_P contains only Π_1-sentences means that Penrose knows that his mind is sound; however, even with this concession, Penrose's argument will fail. In "Mechanism, Truth, and Penrose's New Argument" (2003), Shapiro notes that Penrose's restrictions block the inference to statement (P4) in his argument: *I know M^* is consistent.* [43] This is because Penrose's restrictions prevent him from inferring the soundness of M^*. Recall that M^* is the formal system which outputs a set of sentences, S_{M^*}, such that S_{M^*} contains every sentence output by M plus statement (P1) *My mind is exactly modeled by M* and no other sentences. The statement, "My mind is exactly modeled by M", when uttered by Penrose, is equivalent to the claim that for any L_P-sentence φ, there is a Penrose-proof of φ just in case there is an M-proof of φ. Thus, statement (P1)

[43] Shapiro's formulation of Penrose's argument is slightly different than mine and so his formulation of this problem is slightly different than the one I give here. However, the point is the same.

expresses:

(P1*) $\forall\varphi(Prov_P(\ulcorner\varphi\urcorner) \leftrightarrow Prov_M(\ulcorner\varphi\urcorner))$

This raises a problem for Penroses argument because (P1*) is not a Π_1-sentence. Penrose has restricted the L_P-predicate $U(x)$ such that it is a reliable indicator of truth for only Π_1-sentences. This means that Penrose cannot assert, or suppose, that $U(\ulcorner P1^*\urcorner)$ is true and declare M^* to be sound on that basis. $U(\ulcorner P1^*\urcorner)$ would reliably indicate the truth of (P1*) only if (P1*) were Π_1. (P1*) is not Π_1. Hence, $U(\ulcorner P1^*\urcorner)$ will not reliably indicate the truth of (P1*). This means that Penrose's method for determining the soundness of a system (specifically his mind and M) is not reliable for M^*. Therefore, the inference to statement (P4) in his argument is blocked and so his argument is unsound. [44]

Penrose could respond to Shapiros objection by arguing that he does not need $U(\ulcorner P1^*\urcorner)$ to reliably indicate the truth of (P1*) in order to determine that M^* is sound because he does not need to appeal to $U(\ulcorner P1^*\urcorner)$ in order to make that determination. Penrose only needs the L_P-predicate $U(x)$ to demonstrate that his mind is sound. Penrose knows that all of the sentences which he proves are unassailable (i.e. $\forall\varphi(Prov_P(\ulcorner\varphi\urcorner) \rightarrow U(\ulcorner\varphi\urcorner))$). Since Penrose knows that $(U(\ulcorner\varphi\urcorner) \leftrightarrow \varphi)$ is true for all Π_1-sentences and since Penrose is only able to prove Π_1-sentences, Penrose knows that all of the sentences he proves are true. Hence, Penrose knows that his mind is sound and, by the assumption that M exactly models his mind, Penrose knows that M is sound too. Since Penrose knows that M is sound, he can infer the soundness of M^* directly from the assumption that (P1*) is true. Penrose's argument begins with the assumption

[44] For Shapiro's formulation of this problem see (Shapiro 2003 pp. 31–32).

that (P1*) is true. S_{M^*} contains every sentence output by M plus (P1*) and no other sentences. Penrose knows that every sentence output by M is true (since he knows that M is sound) and, by assumption, Penrose knows that (P1*) is true. Hence, Penrose knows (by assumption and M's soundness) that S_{M^*} contains only true sentences. Therefore, Penrose knows (by assumption and M's soundness) that M^* is sound. Therefore, Penrose does not need $U(\ulcorner P1^* \urcorner)$ to reliably indicate the truth of (P1*) in order to determine that M^* is sound because Penrose does not need to appeal to $U(\ulcorner P1^* \urcorner)$ in order to make that determination.

Even if it is conceded that Penrose knows that the soundness of M^* follows from the assumption that (P1*) is true, Penrose's argument still fails because it is not possible for both statement (P2) and statement (P5) to be true. Recall that in order for statement (P2) to be true, Penrose needs to restrict the kinds of sentences which he can assert to only Π_1-sentences, and he needs to restrict unassailability as being a reliable indicator of truth for only Π_1-sentences. If statement (P2) is true, then at least these two restrictions need to be in place. [45] Statement (P5) is the claim that Penrose is exactly modeled by M^* (when statement (P5) is uttered by Penrose). So, statement (P5) is true just in case:

(P5*) $\quad \forall\varphi((Prov_P(\ulcorner\varphi\urcorner) \leftrightarrow (Prov_{M^*}(\ulcorner\varphi\urcorner)))$

is true. It follows from the definition of M* that $Prov_{M^*}(\ulcorner P1^* \urcorner)$ is true; however, if statement (P2) is true, then $Prov_P(\ulcorner P1^* \urcorner)$ cannot be true because (P1*) is not a Π_1-sentence and Penrose cannot prove (i.e. assert) non-Π_1-sentences. Hence, Penrose must reject either statement (P2), *I know that my*

[45] Recall that if these two restrictions are not in place, then it is not possible for Penrose to consistently know that he is sound.

mind is sound or else statement (P5) *I am exactly modeled by* M^*. If Penrose rejects statement (P2), then his argument collapses immediately. Alternatively, if Penrose rejects statement (P5) then even if he knows that M^* is sound, and thereby knows that M^* is consistent, it will no longer follow that M^* knows that M^* is consistent. Hence, if Penrose rejects statement (P5), then his argument will fail to generate a contradiction from the assumption that he is exactly modeled by M. Since either statement (P2) or (P5) must be false and Penroses argument collapses unless both are true, Penroses argument fails to falsify Mechanism. [46]

3.5 Concluding Remarks on Penrose's Argument

Penrose's argument does not conclusively show that Gödel's theorems are incompatible with Mechanism. There are a number of problems with Penrose's argument, each of which seems to stem from Penrose's claim to know that his mind is sound (i.e. statement (P2)). It was shown that Penrose's method for demonstrating his own soundness implies that he is inconsistent, unless he imposes certain restrictions. However, even with the restrictions, Penrose's "demonstration" of his own soundness remains unconvincing because Penrose does not really *demonstrate* that his mind is sound, he *stipulates* it. Moreover, the restrictions that Penrose requires to make (P2) true block inferences

[46] At this juncture it could be argued that Penrose's ability to demonstrate his own soundness is enough to show, from Gödel's theorems that he cannot be exactly modeled by any formal system. Going this route would require tweaking Penrose's argument slightly, but ultimately it will remain unconvincing (given Penrose's current argument for his own soundness). As was pointed out at the end of §3.3, Penrose does not actually *demonstrate* his own soundness, he merely stipulates it. In order to make a convincing case against Mechanism, Penrose would need to give a better demonstration of his own soundness. As yet, no such demonstration is forthcoming.

to other key premises in his argument. The inference to (P4) *I know that M^* is consistent* (when uttered by Penrose) is blocked because Penroses method for determining soundness is not reliable for M^*. Additionally, it is not possible for both (P2) and (P5) *I am exactly modeled by M^** (when uttered by Penrose) to both be true because if (P2) is true, then Penrose is prohibited from proving non-Π_1-sentences and M^* proves (at least one) non-Π_1-sentence. Yet, if either (P2) or (P5) is rejected, Penrose's argument no longer shows that a contradiction follows from the assumption that he is exactly modeled by some formal system, M.

Neither Lucas's argument nor Penroses argument convincingly demonstrates that Gödel's theorems are incompatible with Mechanism. David Chalmers observes that,

> [T]he deepest flaw [in Gödelean arguments against Mechanism] lies in the assumption that we know that we are sound. All Gödelean arguments appeal to this premise somewhere, but in fact the premise generates a contradiction. Perhaps we are sound, but we cannot know unassailably that we are sound. (1995 § 3.14)

Chalmers seems right, at least in part. It appears that the major weakness of both arguments is an inability to convincingly demonstrate that the human mind is capable of (consistently) proving its own soundness or consistency or the consistency of M. Perhaps what is needed is an argument for the incompatibility of Gödel's theorems with Mechanism which does not appeal to the mind's ability to prove its own soundness or consistency or the consistency of M. Storrs McCall develops such an argument. [47] In the next section I

[47] Note: This is not quite accurate. McCall's argument *does*, in a sense, rely on a consis-

will examine McCalls attempt to show that the human mind can gain certain knowledge from Gödel's incompleteness theorems which no formal system can possess.

4 The McCall Argument

In his paper, "Can a Turing Machine Know that the Gödel Sentence is True?" (1999) McCall presents an innovative argument intended to show that Gödel's theorems are (indirectly) incompatible with Mechanism. [48] He points out that Gödel's incompleteness theorems can be used to demonstrate that truth and provability diverge and that this divergence, and the human mind's awareness of it, points to a fundamental difference between human and machine thinking. He makes two main points. The first is that the human mind is able to know when a (particular) sentence is true and not provable, but no formal system has such knowledge. The second is that no formal system is capable of knowing about the existence of category *true but not provable*. It is this second point which makes McCall's argument so unique. Unlike the other arguments which have been examined so far, McCall's second point does not rely on his being able to demonstrate his own consistency or the consistency of the formal system which is alleged to exactly model his mind in order to

tency proof. He is required to assume the consistency of his mind and of the formal system which exactly models it in order to draw his conclusion. If his claim is that no formal system can prove what *he* can prove, then the consistency of each seems presupposed. Else, each could prove everything.

[48] My use of 'indirectly' in parentheses is intended to make it explicit that that McCall's argument (if successful) shows that Gödel's theorems are *indirectly* incompatible with Mechanism. McCall's argument is supposed to show that the human mind is able to learn something from Gödel's results which no formal system is capable of learning. So, it is not Gödel's theorems *per se* which are incompatible with Mechanism, rather it is what we can learn from them.

draw the conclusion that Mechanism is false. [49] This allows McCall to avoid the primary difficulties associated with the Lucas and Penrose arguments. Still, it is not clear that McCall is successful in his efforts. Below I will show that McCall's argument fails to show that Gödel's theorems are (indirectly) incompatible with Mechanism despite his novel approach.

I have divided this section into three main subsections. In section (4.1), I present McCalls argument. In section (4.2), I discuss McCalls first claim that he, unlike a formal system, is able to know when a (particular) sentence is true but not provable. I argue that McCall does not show that he has such knowledge when the formal system under consideration is the one which is alleged to exactly model his mind. In section (4.3), I explore McCall's second claim and argue that it is possible for a formal system to know about the existence of the category *true but not provable*.

4.1 McCall's Argument

McCalls argument may be put as follows:

(M1) Gödel's incompleteness theorems show that for any formal system, M, if M is consistent, then its Gödel's sentence, G_M, is not provable (in M) but true.

(M2) Gödel's incompleteness theorems show that for any formal system, M, if M is inconsistent, then its Gödel sentence, G_M, is provable (in M) but false.

[49] See Footnote: 47

(M3) So, Gödel's incompleteness theorems show that truth and provability diverge.

(M4) The statement "if M is consistent, then $\neg G_M$ is not provable (in M)" (call it TNP) is true but not provable (in M).

(M5) McCall knows that TNP is *true but not provable (in M)* but M cannot know that.

(M6) McCall knows about the existence of the category *true but not provable* but M cannot.

(M7) Therefore, McCall is not exactly modeled by M.

Statement (M7) is supposed to follow independently from either statement (M5) or (M6). If either claim is true, where M is supposed to exactly model McCall's mind, then his argument succeeds as a refutation of Mechanism. Below I will examine each of these claims in turn and show that both are false where M is supposed to exactly model McCall's mind.

4.2 McCall's First Claim: Statement (M5)

In statement (M5), McCall makes two crucial claims. Firstly, that he knows that TNP is true but not provable (in M) and secondly that M does not know that TNP is true but not provable (in M). With respect to the first claim, McCall thinks that the truth of TNP appears to follow directly from G1. G1 shows that if M is consistent, then G_M is undecidable in M. Hence, M cannot prove $\neg G_M$ (if it is consistent). [50] So the statement, "If M is consistent then

[50] Note: McCall is somewhat mistaken here. G1 implies that $\neg G_M$ is unprovable in M with the stronger assumption that M is ω-consistent. This comes up in § 4.2.1.

$\neg G_M$ is not provable (in M)" (i.e. TNP) must be true. McCall then argues that TNP is unprovable in M, though he does not give a decisive argument to that effect. McCall insists merely that it is unlikely that M could produce a proof of TNP. McCall uses PA to illustrate his point noting,

> [T]he consequent...of $[Con_{PA} \rightarrow \neg Prov_{PA}(\ulcorner Prov_{PA}(\ulcorner G_{PA} \urcorner) \urcorner)]$ is a net strengthening of the consequent...of $[Con_{PA} \rightarrow \neg Prov_{PA}(\ulcorner G_{PA} \urcorner)]$ and it is unlikely that the long and complex proof of $[Con_{PA} \rightarrow \neg Prov_{PA}(\ulcorner G_{PA} \urcorner)]$ in PA could be reworked so as to yield a proof of $[Con_{PA} \rightarrow \neg Prov_{PA}(\ulcorner Prov_{PA}(\ulcorner G_{PA} \urcorner) \urcorner)]$.[51] (1999 p. 529)

Of course, this falls short of a proof that the PA-analog of TNP (i.e. '$Con_{PA} \rightarrow \neg Prov_{PA}(\ulcorner Prov_{PA}(\ulcorner G_{PA} \urcorner) \urcorner)$') is not provable in PA, but I will not harp on that point now. Instead, I will grant that McCall is right that PA cannot, in fact, prove its analog of TNP. McCall goes on to claim that every formal system which extends PA will have a TNP analog of its own. [52] Since it is unlikely that PA can prove its TNP analog, it is also unlikely that any formal system which extends PA will be able to prove *its* TNP analog (provided the formal system is consistent). Hence, if one grants that PA cannot prove its TNP analog, one should also grant that no formal system which extends PA can prove its TNP analog either. Since McCall knows that M will at least extend PA where M is supposed to exactly model his mind, McCall knows that TNP is true but not provable for M where M is supposed to exactly model his mind.

[51] I have changed McCall's notation and omitted some of his abbreviations to better match the notation used throughout this paper as well as to (I hope) better facilitate readability. The content remains the same.

[52] (McCall 1999 p. 531)

McCall thinks that, unlike him, M cannot know that TNP is true but not provable in M (i.e. the second crucial claim of (M5) is true). The reason is rather straight forward. McCall thinks that 'knowability' is identical with 'provability' when it comes to formal systems. He writes, "A [formal system] can *know* what it can *prove*, that is, deduce from axioms using well-defined rules of inference. Its axiomatic database may be large, and its rules of inference may be efficient and powerful. But for a [formal system], *knowability = provability*." (1999 p. 525) Granting him this equivalence, it would seem that M cannot know that TNP is true but not provable in M (if M is consistent). If M were to know that TNP is true but not provable in M, then M would prove that it is not the case that there is an M-proof of TNP. It does not seem that M could know that TNP is also *true* unless M proves TNP. Yet, if M proves TNP then M will prove that there is an M-proof of TNP. M cannot do this if it is consistent. [53] McCall concludes that M cannot know that TNP is true but not provable in M. Hence, statement (M5) in his argument is true and he cannot be exactly modeled by M.

4.2.1 Statement (M5) is False

Contrary to what McCall argues, statement (M5) is false where M is supposed to exactly model McCall's mind. The key to showing that (M5) is false (where M is supposed to exactly model McCall's mind) rests on McCall's assertion that *knowability* is identical with *provability* for formal systems. If McCall's argument is to convincingly show that Gödel's theorems are (indirectly) incom-

[53] The reason for this has to do with Σ_1-completeness. The problem is analogous to the one described on pp. 52–53 (also see footnotes 58 and 60).

patible with Mechanism, then McCall needs to regard *knowability* as identical with *provability* with respect to human subjects also. If McCall claims otherwise (i.e. that *knowability* is not identical with *provability* with respect to human subjects), then his conclusion, (M7) *McCall is not exactly modeled by* M would no longer follow from (M5) (or (M6)). To show that McCall is not exactly modeled by M, McCall must show that he can *prove* something which M cannot (or vice versa I suppose). Since statement (M5) (and (M6)) claim only that McCall has *knowledge* which M cannot have, in order for McCall's conclusion to follow *knowability* and *provability* must be identical for McCall too (at least with respect to the knowledge claims made in (M5) and (M6)). With this in mind, it is possible to show that statement (M5) is false.

Strictly speaking, statement (M5) is false where M is supposed to exactly model McCall's mind because McCall does not *prove* that TNP is true but not provable in M. However, there is another reason to doubt (M5). McCall defends (M5) by arguing that PA cannot prove its analog of TNP and that he knows that the PA analog of TNP is true by G1. McCall then extends this result to apply to all extensions of PA (which will include M where M exactly models his mind). However, McCall's defense of (M5) is problematic. Alexander George and D. J. Velleman (2000) point out that McCall is not actually capable of establishing the PA analog of TNP: *if PA is consistent, then* $\neg G_{PA}$ *is not provable (in PA)* by G1. Rather, by G1 McCall is capable of establishing TNP′: *if PA is* ω*-consistent, then* $\neg G_{PA}$ *is not provable (in PA)*. That $\neg G_{PA}$ is not provable in PA cannot be inferred from the mere consistency of PA alone. The stronger assumption of ω-consistency is needed to show that $\neg G_{PA}$ is not provable in PA. If PA were ω-*in*consistent it could

be possible for $\neg G_{PA}$ (i.e. $\exists x(Prf_{PA}(x, \ulcorner G_{PA} \urcorner))$) to be derived in PA and yet for each number n to fail to be the Gödel number of a PA-proof of G_{PA}. [54] If PA were ω-inconsistent, it would not entail that PA is inconsistent. That is, it is possible for PA to be consistent and ω-inconsistent. [55] Thus, that $\neg G_{PA}$ is not provable in PA cannot be inferred from the mere consistency of PA by G1. Since McCall is only capable of establishing TNP′ and not (the PA analog of) TNP (by G1), his defense of (M5) fails. [56] As (M5) is, strictly speaking, false and McCall's defense of (M5) fails, statement (M5) can be rejected.

4.3 McCall's Second Claim: Statement (M6)

McCall can respond to the criticism that (M5) is false by pointing out that his argument does not need (M5) in order to show that Gödel's theorems are (indirectly) incompatible with Mechanism. Indeed, McCall thinks that the main thrust of his argument comes from (M6), *McCall knows about (the existence of) the category 'true but not provable' but M cannot.* He writes, "Recognition of the existence of such a category therefore marks the difference in principle between human and machine thinking." (1999 p. 529) If McCall is correct, it is (M6) which shows that Mechanism is false. Like statement (M5), statement (M6) is made up of two parts. Firstly, that McCall knows

[54] A theory T, is ω-inconsistent just in case for some open formula, $F(x)$ T proves $\exists x F(x)$, yet for each number, n, T proves $\neg F(n)$. So if PA were ω-inconsistent it could prove $\exists x(Prf_{PA}(x, \ulcorner G_{PA} \urcorner))$ and for each number n, PA could prove $\neg Prf_{PA}(n, \ulcorner G_{PA} \urcorner)$.

[55] This can be illustrated by considering a theory T, which is PA plus $\neg G_{PA}$ as an additional axiom. If PA is consistent, then so is T (but T is ω-inconsistent). If T is not consistent, then PA plus $\neg G_{PA}$ implies a contradiction. So, PA proves G_{PA}, but if PA is consistent, then that is not possible (by G1). Hence, if PA is consistent, then so is T. But T proves $\exists x(Prf_{PA}(x, \ulcorner G_{PA} \urcorner))$ (i.e.$\neg G_{PA}$) and since T has all of PA's axioms it must also prove for each number n, that $\neg Prf_{PA}(n, \ulcorner G_{PA} \urcorner)$. Hence, T is ω-inconsistent and consistent (if PA is consistent). (This illustration is adapted from (Smith 2007 p. 144))

[56] (George & Velleman 2000 pp. 548–549); also see (Tennant 2001)

about the existence of the category *true but not provable* and secondly that no formal system can know about the existence of that category. The first part of statement (M6) is meant to be uncontroversial, given the fact that McCall understands Gödel's incompleteness theorems. Gödel's theorems show that truth and provability diverge and McCalls recognition of this divergence makes him aware of the existence of the category *true but not provable*.

It seems fairly obvious that the human mind knows about the existence of the category *true but not provable*; however, it is perhaps less obvious that M cannot know about the existence of that category. One can construct a predicate within the language of M, call it L, which expresses the concept of 'provability'. If one could also construct an L-predicate which expresses the concept of 'truth', then M could know about the category *true but not provable* by proving that the former (proof) predicate does not apply to (the Gödel number of) some formula, while the latter (truth) predicate does. McCall argues that this is impossible. He claims that Tarski's theorem proves that,

> The notion of "truth" in PA is...on quite a different footing from that of "provability." The latter concept is represented by an open arithmetical formula; no analogous formula expresses (much less represents) the former, thus reinforcing the hypothesis that "truth," though meaningful to humans is a closed book to a [formal system]. (1999 p. 530)

McCall is correct (in a sense). Tarski's theorem demonstrates two things. First that *no formal system (which is consistent, p.r. axiomatized and extends Robinson arithmetic (Q)) can define 'truth' (i.e. a truth predicate) for its own language, and second that no language, L, rich enough to formulate (a theory*

equivalent to) Q can express the property of numbering an L-truth. It seems to me that McCall's appeal is to the first.

If M cannot define 'truth' in the same way that it can define 'provability', then M cannot know about the category of *true but not provable* because M will be unable to prove of any sentence that it is both *true* and *not provable*. Tarski proved that no formal system (which is consistent, p.r. axiomatized and extends Q) can define 'truth' for its own language. Hence, M cannot define 'truth' for its own language. [57] To see why suppose otherwise. Let M be a consistent, p.r. axiomatized formal system which extends Q and which has a language L and suppose that M *can* define 'truth' for L. So, there is a predicate of L, call it $Tru(x)$ such that for any sentence of L, φ:

$$M \vdash \varphi \leftrightarrow Tru(\ulcorner \varphi \urcorner)$$

M is strong enough to prove the diagonalization lemma, so there is a fixed point, λ, for $\neg Tru(x)$ such that

$$M \vdash \lambda \leftrightarrow \neg Tru(\ulcorner \lambda \urcorner)$$

Since λ is a sentence of L and since, by supposition, M proves $\varphi \leftrightarrow Tru(\ulcorner \varphi \urcorner)$ for all sentences of L:

$$M \vdash \lambda \leftrightarrow Tru(\ulcorner \lambda \urcorner)$$

It follows that M cannot be consistent, contra the initial supposition. Ergo, M cannot define 'truth' for its own language. Therefore, McCall contends, it is not possible for M to know about the category *true but not provable*.

[57] I take it that McCall thinks that if M is supposed to exactly model his mind, then M will be such that Tarski's theorem will apply to it. Of course, as has been discussed before, the mechanist has the option of saying that McCall's mind is inconsistent and so the formal system which exactly models his mind is inconsistent too.

There is an additional problem, which McCall does not mention. It seems that even if M could define the concept of 'truth' (for a language other than its own), it would not be able to consistently know about the concept of *true but not provable* (that is, 'true but not M-provable') provided M is Σ_1-complete. [58] Presumably, M would be Σ_1-complete if it is supposed to exactly model McCall's mind because M should, at least extend PA. To see why this leads to problems suppose M can define 'truth' for a language L' (i.e. a language other than its own language, L), and that M knows about the concept 'true but not (M-)provable'. So, for all L'-sentences φ, M proves $\varphi \leftrightarrow Tru(\ulcorner\varphi\urcorner)$ and of some L'-sentence ψ, M proves that:

(i) $Tru(\ulcorner\psi\urcorner) \wedge \neg Prov_M(\ulcorner\psi\urcorner)$[59]

That is, that ψ is true and not provable (in M). Since M proves (i), M also proves each conjunct of (i). So M proves:

$$Tru(\ulcorner\psi\urcorner)$$

And

$$\neg Prov_M(\ulcorner\psi\urcorner)$$

Since ψ is an L'-sentence and M proves $\varphi \leftrightarrow Tru(\ulcorner\varphi\urcorner)$ for all L'-sentences φ, M proves:

$$\psi \leftrightarrow Tru(\ulcorner\psi\urcorner)$$

[58] M is Σ_1-complete just in case for any Σ_1-sentence φ, if φ is true, then M proves φ. A Σ_1-sentence is (or is logically equivalent to) a sentence which begins with one or more unbounded existential quantifiers (followed by a Δ_0 wff).

[59] Given McCall's stipulation that *knowability* is identical with *provability* for M, if M is to know about the category 'true but not provable (in M)', McCall thinks that M needs to prove (i) or something equivalent.

Hence, M proves ψ. Since M proves ψ, there is an M-proof of ψ and so M will prove:

$$Prov_M(\ulcorner\psi\urcorner)$$[60]

Hence, M proves a contradiction and so, even if M could define 'truth' for L', M cannot both be consistent and know about the concept 'true but not (M-)provable'. McCall concludes that M is unable to know about the category of 'true but not provable', and so there is a fundamental difference between M and McCall. Namely, exactly what statement (M6) expresses: *McCall knows about (the existence of) the category 'true but not provable' but M cannot.* Therefore, McCall's argument shows that Gödel's theorems are (indirectly) incompatible with Mechanism.

4.3.1 Statement (M6) is False

McCall's arguments do not show that it is impossible for M to know about the existence of the category *true but not provable.* McCall is right that M cannot define 'truth' *for its own language* without violating Tarski's theorem and it is not possible for M to consistently know about the category 'true but not *(M-)provable*' (provided M is Σ_1-complete). However, neither of those facts imply that M is incapable of knowing about the category *true but not provable* altogether. McCall seems to ignore the possibility that M can know about the existence of the category *true but not provable* provided it is doing so with respect to a language (and formal system) less rich than its own (and other than itself). To illustrate, let M's language, L include a language L^*. Let the

[60] This is because M is Σ_1-complete and $Prov_M(\ulcorner\psi\urcorner)$ is a true Σ_1-sentence (it expresses: $\exists x(Prf_M(x, \ulcorner\psi\urcorner))$).

open L well formed formula $Tru(x)$ be a formal truth-predicate for L^* such that for every L^*-sentence ψ, $Tru(\ulcorner\psi\urcorner) \leftrightarrow \psi$ is true. Let M be a truth-theory for L^* such that M proves $Tru(\ulcorner\psi\urcorner) \leftrightarrow \psi$ for every L^*-sentence ψ. Let M_1 be a consistent formal system with language, L^*, and let M be richer than M_1. [61] With this in mind, it is entirely possible that for some L^*-sentence, λ, M proves:

(i′) $Tru(\ulcorner\lambda\urcorner) \wedge \neg Prov_{M_1}(\ulcorner\lambda\urcorner)$

That is, it is possible for M to prove that λ is true but not provable in M_1 without resulting in contradiction. (i′) does not represent a violation of Tarski's theorem nor does (i′) lead to contradiction based on M's Σ_1-completeness. Hence, it is possible for M to know about the existence of the category 'true but not provable' with respect to a language (and formal system) less rich than its own (and other than itself). Therefore, statement (M6) is, strictly speaking, false.

McCall may respond by insisting that it is not the knowledge of the existence of a narrow category of *true but not provable* which he has in mind. That is, showing that (M6) is false requires more than showing that it is possible for M to know about the existence of the category *true but not provable* with respect to a language (and formal system) less rich than its own (and other than itself). In order to show that statement (M6) is false it must be demonstrated that M knows about the existence of the category *true but not provable* with respect to *every* language and *every* (consistent) formal system. After all, it seems that the human mind recognizes that this is a category which

[61] M must be rich enough to prove that λ is not provable in M_1 and rich enough to prove λ.

exists with respect to every language and for every (consistent) formal system. It is McCall's ability to recognize the existence of the category in this broad sense which marks the difference between his mind and M. Thus, McCall is not really claiming (M6) but (M6*) *McCall knows that the category true but not provable exists for every language and every (consistent) formal system (including for M where M exactly models his mind), but M cannot.* (M6*) is not shown to be false by M's ability to prove (i$'$). Moreover, since McCall's conclusion (M7) follows from (M6*), (M7) stands, *McCall is not exactly modeled by M.* Therefore, McCall's argument shows that Gödel's theorems are (indirectly) incompatible with Mechanism.

If McCall changes (M6) to (M6*) his argument still fails to convincingly demonstrate that Gödel's theorems are (indirectly) incompatible with Mechanism. It is not clear that (M6*) is true. It seems plausible that McCall might "know" that the category *true but not provable* exists for every formal system in some kind of semantic sense, but that is not what McCall needs. To show that (M6*) is true, McCall must prove that there is a sentence in M's language such that it is *true* but *not provable* in M where M is supposed to exactly model McCall's mind. McCall does not do this. Moreover, if Mechanism is true, then Tarski's theorem holds for McCall's mind (if he is consistent). Thus, if Mechanism is true, it is impossible for McCall to prove that there is a sentence in M's language such that it is *true* but *not provable* in M where M is supposed to exactly model his mind. Hence, changing (M6) to (M6*) seems to be unjustified at best and question begging at worst. Therefore, if McCall changes (M6) to (M6*) his argument still fails to convincingly demonstrate that Gödel's theorems are (indirectly) incompatible with Mechanism.

4.4 Concluding Remarks on McCall's Argument

McCall's argument fails to show that Gödel's incompleteness theorems are (indirectly) incompatible with Mechanism. Both of the major claims McCall makes in support of his conclusion that he is not exactly modeled by M, (M5) and (M6), turn out to be false. Statement (M5) is false because McCall does not actually *prove* that TNP is true but not provable in M where M is supposed to exactly model his mind. Moreover, McCall's argument that TNP is true but not provable in M is entirely unconvincing. Statement (M6) is false because it *is* possible for M to know about the existence of the category *true but not provable* with respect to a language (and formal system) less rich than its own (and other than itself). Hence, McCall's argument fails to show that he is not exactly modeled by M. Therefore, McCall's argument fails to show that Gödel's incompleteness theorems are (indirectly) incompatible with Mechanism.

5 Conclusions

Each of the three arguments which I have examined has failed to demonstrate that Gödel's incompleteness theorems are incompatible (directly or indirectly) with Mechanism. Lucas's approach fails because he is unable to demonstrate that he can consistently formally (Lucas-)prove Con_M or make use of a formal (non-Lucas-)proof of Con_M in a way that M cannot. Penrose attempts to overcome these difficulties by arguing that he is able to deduce a contradiction from the mere assumption that he is modeled by some formal system M. Ultimately, however, Penrose's attempt fails. The success of Penrose's argu-

ment relies crucially on his ability to demonstrate his own soundness. Penroses method for demonstrating his own soundness either leads to inconsistency, or else blocks the inferences to essential premises in his argument, or can be rejected as unconvincing because Penrose merely stipulates his own soundness, he does not prove it. McCall is able to (in part) overcome the major difficulties associated with the Penrose and Lucas arguments surrounding consistency and soundness proofs. McCall argues that Gödel's incompleteness theorems give him knowledge of the existence of the category *true but not provable*. He argues that it is impossible for *any* formal system (which extends PA) to have such knowledge; yet, it was shown that it *is* possible for a formal system to know about the category *true but not provable* with respect to a language (and formal system) less rich than its own (and other than itself).

The above three attempts to demonstrate that Gödel's theorems are incompatible with Mechanism have each failed, but I think something more general can be said. Each of the arguments takes a different approach to the issue, yet they all share a common feature. They each attempt to demonstrate that given Gödel's results, the human mind is capable of proving *some* sentence which no formal system M is capable of proving. Indeed, it seems that, minimally, any argument which purports to show that Gödel's theorems are incompatible with Mechanism (directly or indirectly) must demonstrate precisely that. The trouble is that any such argument (which could currently exist) will likely be too imprecise to yield any real or definitive conclusions in this matter.

Any argument which purports to show that Gödel's theorems are incompatible with Mechanism, and which could currently exist, will likely be too imprecise to yield any real or definitive conclusions. The arguments examined

above serve as a good illustration of why this is the case. Each author made use of sentences like 'Con_M' and 'G_M', but how are these sentences to be understood? The authors seem only to define these sentences in terms of what they indirectly state (and/or for being like their analogs in defined theories like PA). Namely, *there is no n which is the Gödel's number of an M-proof of a contradiction or of* G_M. But what is an 'M-proof'? What are M's axioms? What are M's inference rules? Without answering questions like these, sentences like 'Con_M' and 'G_M' remain, essentially, undefined. In his criticism of Lucas's argument Benacerraf writes:

> If given a black box and told not to peek inside, then what reason is there to suppose that Lucas or I can determine its program by watching its output? But I must be able to determine its program (if that makes sense) if I am to carry out Gödel's argument in connection with it...If the machine is not designated in such a way that there is an effective procedure for recovering the machine's program from the designation, one may well know that one is presented with a machine but yet be unable to do anything about finding the Gödel's sentence for it. (1967 p. 28)

He goes on to say:

> *In a relevant sense, if I am a Turing machine, then perhaps I cannot ascertain which one.* In the absence of such knowledge, I can cheerfully go around 'proving' my own consistency, but not in an arithmetic waynot using my own proof predicate. (1967 p. 29)

Benacerraf's point is that without the benefit of clear definitions, in partic-

ular of 'provability-in-M' or 'provability-in-one's mind', it is not possible to definitively show that Gödel's incompleteness theorems are incompatible with Mechanism. Even if a Gödelean argument against Mechanism does not contain any other problems, as long as it appeals to undefined provability predicates, the argument will be too hypothetical to yield any real results.

Showing that there is a *real* incompatibility between Gödel's theorems and Mechanism requires using defined notions of provability. One's claiming to be able to formally prove Con_M, for instance, will not be very convincing, let alone have any obvious implications, unless one can define what 'Con_M' means. Therefore, until such time as someone can define a provability predicate for the human mind *or* for the formal system alleged to exactly model the human mind (which may be decades, centuries, or perhaps never), it will remain improbable that Gödel's theorems can be definitively shown to be incompatible with Mechanism.

References

Benacerraf, P. (1967). "God, The Devil, and Gödel, " *The Monist*, 51: 9–32.

Boolos, G. S. (1968). "Minds, Machines and Gödel by J. R. Lucas; God, the Devil and Gödel by Paul Benacerraf," *The Journal of Symbolic Logic*, 33: 613–615.

Chalmers, David J. (1995). "Minds, Machines, And Mathematics: A Review of Shadows of the Mind by Roger Penrose," *Psyche*, 2(9): http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html.

Franzén, T. (2005). *Gödel's Theorem: An Incomplete Guide To Its Use And Abuse*, Natick: A K Peters.

George, A. & Velleman, D. J. (2000). "Leveling the Playing Field between Mind and Machine: A Reply to McCall, " *The Journal of Philosophy*, 97: 456–461.

Lewis, D. (1979). "Lucas against Mechanism II," *Canadian Journal of Philosophy*, 9: 373–376.

Lindström, P. (2006). "Remarks on Penrose's "New Argument", " *Journal of Philosophical Logic*, 35: 231–237.

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

Lucas, J. R. (1968). "Satan Stultified: A Rejoinder to Paul Benacerraf, " *The Monist*, 52: 145–158.

McCall, S. (1999). "Can a Turing Machine Know that the Gödel Sentence is True? ," *The Journal of Philosophy*, 96: 525–532.

Penrose, R. (1996). *Shadows of the Mind*, New York: Oxford University Press.

Penrose, R. (1996)-B. "Beyond the Doubting of a Shadow: A Reply to Commentaries on Shadows of the Mind, " *Psyche*, 2(23): http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html

Shapiro, S. (1998). "Incompleteness, Mechanism, and Optimism, " *The Bulletin of Symbolic Logic*, 4: 273–302.

Shapiro, S. (2003). "Mechanism, Truth, and Penrose's New Argument," *Journal of Philosophical Logic*, 32: 19–42.

Smith, P. (2007). *An Introduction to Gödel's Theorems*, New York: Cambridge University Press.

Tennant, N. (2001). "On Turing Machines Knowing Their Own Gödel-Sentences," *Philosophia Mathematica*, 9: 72–79.