The Lucas-Penrose Argument about Gödel's Theorem

In 1961, J.R. Lucas published "Minds, Machines and Gödel," in which he formulated a controversial anti-mechanism argument. The argument claims that Gödel's first incompleteness theorem shows that the human mind is not a Turing machine, that is, a computer. The argument has generated a great deal of discussion since then. The influential Computational Theory of Mind, which claims that the human mind is a computer, is false if Lucas's argument succeeds. Furthermore, if Lucas's argument is correct, then "strong artificial intelligence," the view that it is possible at least in principle to construct a machine that has the same cognitive abilities as humans, is false. However, numerous objections to Lucas's argument have been presented. Some of these objections involve the consistency or inconsistency of the human mind; if we cannot establish that human minds are consistent, or if we can establish that they are in fact inconsistent, then Lucas's argument fails (for reasons made clear below). Others object to various idealizations that Lucas's argument makes. Still others find some other fault with the argument. Lucas's argument was rejuvenated when the physicist R. Penrose formulated and defended a version of it in two books, 1989's *The Emperor's New Mind* and 1994's *Shadows of the Mind*. Although there are similarities between Lucas's and Penrose's arguments, there are also some important differences. Penrose argues that the Gödelian argument implies a number of claims concerning consciousness and quantum physics; for example, consciousness must arise from quantum processes and it might take a revolution in physics for us to obtain a scientific explanation of consciousness. There have also been objections raised to Penrose's argument and the various claims he infers from it: some question the anti-mechanism argument itself, some question whether it entails the claims about consciousness and physics that he thinks it does, while others question his claims about consciousness and physics, apart from his anti-mechanism argument.

Section one discusses Lucas's version of the argument. Numerous objections to the argument – along with Lucas's responses to these objections – are discussed in section two. Penrose's version of the argument, his claims about consciousness and quantum physics, and various objections that are specific to Penrose's claims are discussed in section three. Section four

briefly addresses the question, "What did Gödel himself think that his theorem implied about the human mind?" Finally, section five mentions two other anti-mechanism arguments.

Table of Contents

1. Lucas's Original Version of the Argument

Gödel's (1931) first incompleteness theorem proves that any consistent formal system in which a "moderate amount of number theory" can be proven will be incomplete, that is, there will be at least one true mathematical claim that cannot be proven within the system (Wang 1981: 19). The claim in question is often referred to as the "Gödel sentence." The Gödel sentence asserts of itself: "I am not provable in *S*," where "*S*" is the relevant formal system. Suppose that the Gödel sentence *can be proven* in *S*. If so, then by soundness the sentence is true in *S*. But the sentence claims that it is not provable, so it must be that we *cannot prove* it in *S*. The assumption that the Gödel sentence is provable in *S* leads to contradiction, so if *S* is consistent, it must be that the Gödel sentence is unprovable in *S*, and therefore true, because the sentence claims that it is not provable. In other words, if consistent, *S* is incomplete, as there is a true mathematical claim that cannot be proven in *S*. For an introduction to Gödel's theorem, see Nagel and Newman (1958).

Gödel's proof is at the core of Lucas's (1961) argument, which is roughly the following. Consider a machine constructed to produce theorems of arithmetic. Lucas argues that the operations of this machine are analogous to a formal system. To explain, "if there are only a definite number of types of operation and initial assumptions built into the [machine], we can

represent them all by suitable symbols written down on paper" (Lucas 1961: 115). That is, we can associate specific symbols with specific states of the machine, and we can associate "rules of inference" with the operations of the machine that cause it to go from one state to another. In effect, "given enough time, paper, and patience, [we could] write down an analogue of the machine's operations," and "this analogue would in fact be a formal proof" (ibid). So essentially, the arithmetical claims that the machine will produce as output, that is, the claims the machine proves to be true, will "correspond to the theorems that can be proved in the corresponding formal system" (ibid). Now suppose that we construct the Gödel sentence for this formal system. Since the Gödel sentence cannot be proven in the system, the machine will be unable to produce this sentence as a truth of arithmetic. However, a human can look and see that the Gödel sentence is true. In other words, there is at least one thing that a human mind can do that no machine can. Therefore, "a machine cannot be a complete and adequate model of the mind" (Lucas 1961: 113). In short, the human mind is not a machine.

Here is how Lucas (1990: paragraph 3) describes the argument:

> I do not offer a simple knock-down proof that minds are inherently better than machines, but a schema for constructing a *dis*proof of any plausible mechanist thesis that might be proposed. The disproof depends on the particular mechanist thesis being maintained, and does not claim to show that the mind is uniformly better than the purported mechanist representation of it, but only that it is one respect better and therefore different. That is enough to refute that particular mechanist thesis.

Further, Lucas (ibid) believes that a variant of his argument can be formulated to refute any future mechanist thesis. To explain, Lucas seems to envision the following scenario: a mechanist formulates a particular mechanistic thesis by claiming, for example, that the human mind is a Turing machine with a given formal specification *S*. Lucas then refutes this thesis by producing *S*'s Gödel sentence, which we can see is true, but the Turing machine cannot. Then, a mechanist puts forth a different thesis by claiming, for example, that the human mind is a Turing machine with formal specification *S'*. But then Lucas produces the Gödel sentence for *S'*, and so on, until, presumably, the mechanist simply gives up.

One who has not studied Gödel's theorem in detail might be wondering: why can't we simply add the Gödel sentence to the list of theorems a given machine "knows" thereby giving the machine the ability Lucas claims it does not have? In Lucas's argument, we consider some particular Turing machine specification *S*, and then we note that "*S*-machines" (that is, those machines that have formal specification *S*) cannot see the truth of the Gödel sentence while we can, so human minds cannot be *S*-machines, at least. But why can't we simply add the Gödel sentence to the list of theorems that *S*-machines can produce? Doing so will presumably give

the machines in question the ability that allegedly separates them from human minds, and Lucas's argument falters. The problem with this response is that even if we add the Gödel sentence to *S*-machines, thereby producing Turing machines that can produce the initial Gödel sentence as a truth of arithmetic, Lucas can simply produce a new Gödel sentence for these updated machines, one which allegedly we can see is true but the new machines cannot, and so on ad infinitum. In sum, as Lucas (1990: paragraph 9) states, "It is very natural...to respond by including the Gödelian sentence in the machine, but of course that makes the machine a different machine with a different Gödelian sentence all of its own." This issue is discussed further below.

One reason Lucas's argument has received so much attention is that if the argument succeeds, the widely influential Computational Theory of Mind is false. Likewise, if the argument succeeds, then "strong artificial intelligence" is false; it is impossible to construct a machine that can perfectly mimic our cognitive abilities. But there are further implications; for example, a view in philosophy of mind known as Turing machine functionalism claims that the human mind is a Turing machine, and of course, if Lucas is right, this form of functionalism is false. (For more on Turing machine functionalism, see Putnam (1960)). So clearly there is much at stake.

2. Some Possible Objections to Lucas

Lucas's argument has been, and still is, very controversial. Some objections to the argument involve consistency; if we cannot establish our own consistency, or if we are in fact inconsistent, then Lucas's argument fails (for reasons made clear below). Furthermore, some have objected that the algorithm the human mind follows is so complex we might be forever unable to formulate our own Gödel sentence; if so, then maybe we cannot see the truth of our own Gödel sentence and therefore we might not be different from machines after all. Others object to various idealizations that Lucas's argument makes. Still others find some other fault with the argument. In this section, some of the more notable objections to Lucas's argument are discussed.

a. Consistency

Lucas's argument faces a number of objections involving the issue of consistency; there are two related though distinct lines of argument on this issue. First, some claim that we cannot establish our own consistency, whether we are consistent or not. Second, some claim that we are in fact inconsistent. The success of either of these objections would be sufficient to defeat Lucas's argument. But first, to see why these objections (if successful) would defeat Lucas's

argument, recall that Gödel's first incompleteness theorem states that *if* a formal system (in which we can prove a suitable amount of number theory) is consistent, the Gödel sentence is true but unprovable in the system. That is, the Gödel sentence will be true and unprovable only in consistent systems. In an inconsistent system, one can prove any claim whatsoever because in classical logic, any and all claims follow from a contradiction; that is, an inconsistent system will not be incomplete. Now, suppose that a mechanist claims that we are Turing machines with formal specification *S*, and this formal specification is inconsistent (so the mechanist is essentially claiming that we are inconsistent). Lucas's argument simply does not apply in such a situation; his argument cannot defeat this mechanist. Lucas claims that any machine will be such that there is a claim that is true but unprovable for the machine, and since we can see the truth of the claim but the machine cannot, we are not machines. But if the machine in question is inconsistent, the machine will be able to prove the Gödel sentence, and so will not suffer from the deficiency that Lucas uses to separate machines from us. In short, for Lucas's argument to succeed, human minds must be consistent.

Consequently, if one claims that we cannot establish our own consistency, this is tantamount to claiming that we cannot establish the truth of Lucas's conclusion. Furthermore, there are some good reasons for thinking that even if we are consistent, we cannot establish this. For example, Gödel's second incompleteness theorem, which quickly follows from his first theorem, claims that one cannot prove the consistency of a formal system *S* from within the system itself, so, if we are formal systems, we cannot establish our own consistency. In other words, a mechanist can avoid Lucas's argument by simply claiming that we are formal systems and therefore, in accordance with Gödel's second theorem, cannot establish our own consistency. Many have made this objection to Lucas's argument over the years; in fact, Lucas discusses this objection in his original (1961) and attributes it to Rogers (1957) and Putnam. Putnam made the objection in a conversation with Lucas even before Lucas's (1961) (see also Putnam (1960)). Likewise, Hutton (1976) argues from various considerations drawn from Probability Theory to the conclusion that we cannot assert our own consistency. For example, Hutton claims that the probability that we are inconsistent is above zero, and that if we claim that we are consistent, this "is a claim to infallibility which is insensitive to counter-arguments to the point of irrationality" (Lucas 1976: 145). In sum, for Lucas's argument to succeed, we must be assured that humans are consistent, but various considerations, including Gödel's second theorem, imply that we can never establish our own consistency, even if we are consistent.

Another possible response to Lucas is simply to claim that humans are in fact inconsistent Turing machines. Whereas the objection above claimed that we can never establish our own consistency (and so cannot apply Gödel's first theorem to our own minds with complete confidence), this new response simply outright denies that we are consistent. If humans are

inconsistent, then we might be equivalent to inconsistent Turing machines, that is, we might be Turing machines. In short, Lucas concludes that since we can see the truth of the Gödel sentence, we cannot be Turing machines, but perhaps the most we can conclude from Lucas's argument is that either we are not Turing machines or we are inconsistent Turing machines. This objection has also been made many times over the years; Lucas (1961) considers this objection too in his original article and claims that Putnam also made this objection to him in conversation.

So, we see two possible responses to Lucas: (1) we cannot establish our own consistency, whether we are consistent or not, and (2) we are in fact inconsistent. However, Lucas has offered numerous responses to these objections. For example, Lucas thinks it is unlikely that an inconsistent machine could be an adequate representation of a mind. He (1961: 121) grants that humans are sometimes inconsistent, but claims that "it does not follow that we are tantamount to inconsistent systems," as "our inconsistencies are mistakes rather than set policies." When we notice an inconsistency within ourselves, we generally "eschew" it, whereas "if we really were inconsistent machines, we should remain content with our inconsistencies, and would happily affirm both halves of a contradiction" (ibid). In effect, we are not inconsistent machines even though we are sometimes inconsistent; we are fallible but not systematically inconsistent. Furthermore, if we were inconsistent machines, we would potentially endorse any proposition whatsoever (ibid). As mentioned above, one can prove any claim whatsoever from a contradiction, so if we are inconsistent Turing machines, we would potentially believe anything. But we do not generally believe any claim whatsoever (for example, we do not believe that we live on Mars), so it appears we are not inconsistent Turing machines. One possible counter to Lucas is to claim that we are inconsistent Turing machines that reason in accordance with some form of paraconsistent logic (in paraconsistent logic, the inference from a contradiction to any claim whatsoever is blocked); if so, this explains why we do not endorse any claim whatsoever given our inconsistency (see Priest (2003) for more on paraconsistent logic). One could also argue that perhaps the inconsistency in question is hidden, buried deep within our belief system; if we are not aware of the inconsistency, then perhaps we cannot use the inconsistency to infer anything at all (Lucas himself mentions this possibility in his (1990)).

Lucas also argues that even if we cannot prove the consistency of a system from within the system itself, as Gödel's second theorem demonstrates, there might be other ways to determine if a given system is consistent or not. Lucas (1990) points out that there are finitary consistency proofs for both the propositional calculus and the first-order predicate calculus, and there is also Gentzen's proof of the consistency of Elementary Number Theory. Discussing Gentzen's proof in more detail, Lucas (1996) argues that while Gödel's second theorem

demonstrated that we cannot prove the consistency of a system from *within* the system itself, it might be that we can prove that a system is consistent with considerations drawn from *outside* the system. One very serious problem with Lucas's response here, as Lucas (ibid) himself notes, is that the wider considerations that such a proof uses must be consistent too, and this can be questioned. Another possible response is the following: maybe we can "step outside" of, say, Peano arithmetic and argue that Peano arithmetic is consistent by appealing to considerations that are outside of Peano arithmetic; however, it isn't clear that we can "step outside" of ourselves to show that we are consistent.

Lucas (1976: 147) also makes the following "Kantian" point:

> [perhaps] we must assume our own consistency, if thought is to be possible at all. It is, perhaps like the uniformity of nature, not something to be established at the end of a careful chain of argument, but rather a necessary assumption we must make if we are to start on any thinking at all.

A possible reply is that assuming we are consistent (because this assumption is a necessary precondition for thought) and our actually being consistent are two different things, and even if we must assume that we are consistent to get thought off of the ground, we might be inconsistent nevertheless. Finally, Wright (1995) has argued that an intuitionist, at least, who advances Lucas's argument, can overcome the worry over our consistency.

b. Benacerraf's Criticism

Benacerraf (1967) makes a well-known criticism of Lucas's argument. He points out that it is not easy to construct a Gödel sentence and that in order to construct a Gödel sentence for any given formal system one must have a solid understanding of the algorithm at work in the system. Further, the formal system the human mind might implement is likely to be extremely complex, so complex, in fact, that we might never obtain the insight into its character needed to construct our version of the Gödel sentence. In other words, we understand some formal systems, such as the one used in Russell and Whitehead's (1910) *Principia*, well enough to construct and see the truth of the Gödel sentence for these systems, but this does not entail that we can construct and see the truth of our own Gödel sentence. If we cannot, then perhaps we are not different from machines after all; we might be very complicated Turing machines, but Turing machines nevertheless. To rephrase this objection, suppose that a mechanist produces a complex formal system *S* and claims that human minds are *S*. Of course, Lucas will then try to produce the Gödel sentence for *S* to show that we are not *S*. But *S* is extremely complicated, so complicated that Lucas cannot produce *S*'s Gödel sentence, and so cannot disprove this

particular mechanistic thesis. In sum, according to Benacerraf, the most we can infer from Lucas's argument is a disjunction: "*either* no (formal system) encodes all human arithmetical capacity – the Lucas-Penrose thought – *or* any system which does has no axiomatic specification which human beings can comprehend" (Wright, 1995, 87). One response Lucas (1996) makes is that he [Lucas] could be helped in the effort to produce the Gödel sentence for any given formal system/machine. Other mathematicians could help and so could computers. In short, at least according to Lucas, it might be difficult, but it seems that we could, at least *in principle*, determine what the Gödelian formula is for any given system.

c. The Whiteley Sentence

Whiteley (1962) responded to Lucas by arguing that humans have similar limitations to the one that Lucas's argument attributes to machines; if so, then perhaps we are not different from machines after all. Consider, for example, the "Whiteley sentence," that is, "Lucas cannot consistently assert this formula." If this sentence is true, then it must be that asserting the sentence makes Lucas inconsistent. So, either Lucas is inconsistent or he cannot utter the sentence on pain of inconsistency, in which case the sentence is true and so Lucas is incomplete. Hofstadter (1981) also argues against Lucas along these lines, claiming that we would not even believe the Whiteley sentence, while Martin and Engleman (1990) defend Lucas on this point by arguing against Hofstadter (1981).

d. Issues Involving "Idealizations"

A number of objections to Lucas's argument involve various "idealizations" that the argument makes (or at least allegedly makes). Lucas's argument sets up a hypothetical scenario involving a mind and a machine, "but it is an idealized mind and an idealized machine," neither of which are subject to limitations arising from, say, human mortality or the inability of some humans to understand Gödel's theorem, and some believe that once these idealizations are rejected, Lucas's argument falters (Lucas 1990: paragraph 6). Several specific instances of this line of argument are considered in successive paragraphs.

Boyer (1983) notes that the output of any human mind is finite. Since it is finite, it could be programmed into and therefore simulated by a machine. In other words, once we stop ignoring human finitude, that is, once we reject one of the idealizations in Lucas's argument, we are not different from machines after all. Lucas (1990: paragraph 8) thinks this objection misses the point: "What is in issue is whether a computer can copy a living me, when I have not yet done all that I shall do, and can do many different things. It is a question of potentiality rather than actually that is in issue." Lucas's point seems to be that what is really at issue is

what can be done by a human and a machine *in principle*; if, in principle, the human mind can do something that a machine cannot, then the human mind is not a machine, even if it just so happens that any particular human mind could be modeled by a machine as a result of human finitude.

Lucas (1990: paragraph 9) remarks, "although some degree of idealization seems allowable in considering a mind untrammeled by mortality…, doubts remain about how far into the infinite it is permissible to stray." Recall the possible objection discussed above (in section 1) in which the mechanist, when faced with Lucas's argument, responds by simply producing a new machine that is just like the last except it contains the Gödel sentence as a theorem. As Lucas points out, this will simply produce a new machine that has a different Gödel sentence, and this can go on forever. Some might dispute this point though. For example, some mechanists might try "adding a Gödelizing operator, which gives, in effect a whole denumerable infinity of Gödelian sentences" (Lucas 1990: paragraph 9). That is, some might try to give a machine a method to construct an infinite number of Gödel sentences; if this can be done, then perhaps any Gödel sentence whatsoever can be produced by the machine. Lucas (1990) argues that this is not the case, however; a machine with such an operator will have its own Gödel sentence, one that is not on the initial list produced by the operator. This might appear impossible: how, if the initial list is infinite, can there be an additional Gödel sentence that is not on the list? It is not impossible though: the move from the initial infinite list of Gödel sentences to the additional Gödel sentence will simply be a move into the "transfinite," a higher level of infinity than that of the initial list. It is widely accepted in mathematics, and has been for quite some time, that there are different levels of infinity.

Coder (1969) argues that Lucas has an overly idealized view of the mathematical abilities of many people; to be specific, Coder thinks that Lucas overestimates the degree to which many people can understand Gödel's theorem and this somehow creates a problem for Lucas's argument. Coder holds that since many people cannot understand Gödel's theorem, all Lucas has shown is that a handful of competent mathematical logicians are not machines (the idea is that Lucas's argument only shows that those who can produce and see the truth of the Gödel sentence are not machines, but not everyone can do this). Lucas (1970a) responds by claiming, for example, that the only difference between those who can understand Gödel's theorem and those who cannot is that, in the case of the former, it is more obvious that they are not machines; it isn't, say, that some people are machines and others are not.

Dennett (1972) has claimed there is something odd about Lucas's argument insofar as it seems to treat humans as creatures that simply wander around asserting truths of first-order arithmetic. Dennett (1972: 530) remarks,

> Men do not sit around uttering theorems in a uniform vocabulary, but say things in earnest and jest, makes slips of the tongue, speak several languages..., and – most troublesome for this account – utter all kinds of nonsense and contradictions....

Lucas's (1990: paragraph 7) response is that these differences between humans and machines that Dennett points to are sufficient for some philosophers to reject mechanism, and that he [Lucas] is simply giving mechanism the benefit of the doubt by assuming that they can explain these differences. Furthermore, humans can, and some actually do, produce theorems of elementary number theory as output, so any machine that cannot produce all of these theorems cannot be an adequate model of the human mind.

e. Lewis's Objection

Lewis (1969) has also formulated an objection to Lucas's argument:

> Lewis argues that I [that is, Lucas] have established that there is a certain *Lucas arithmetic* which is clearly true and cannot be the output of some Turing machine. If I could produce the whole of Lucas arithmetic, then I would certainly not be a Turing machine. But there is no reason to suppose that I am able in general to verify theoremhood in Lucas arithmetic (Lucas 1970: 149).

To clarify, "Peano arithmetic" is the arithmetic that machines can produce and "Lucas arithmetic" is the arithmetic that humans can produce, and Lucas arithmetic will contain Gödel sentences while Peano arithmetic will not, so humans are not machines, at least according to Lucas's argument. But Lewis (1969) claims that Lucas has not shown us that he (or anyone else, for that matter) can in fact produce Lucas arithmetic in its entirety, which he must do if his argument is to succeed, so Lucas's argument is incomplete. Lucas responds that he does not need to produce Lucas arithmetic in its entirety for his argument to succeed. All he needs to do to disprove mechanism is produce a *single* theorem that a human can see is true but a machine cannot; this is sufficient. Lucas (1970: 149) holds that "what I have to do is to show that a mind can produce not the whole of Lucas arithmetic, but only a small, relevant part. And this I think I can show, thanks to Gödel's theorem."

3. Penrose's New Version of the Argument

Penrose has formulated and defended versions of the Gödelian argument in two books, 1989's *The Emperor's New Mind* and 1994's *Shadows of the Mind*. Since the latter is at least in part an attempt to improve upon the former, this discussion will focus on the latter. Penrose's (1994) consists of two main parts: (a) a Gödelian argument to show that humans minds are

non-computable and (b) an attempt to infer a number of claims involving consciousness and physics from (a). (a) and (b) are discussed in successive sections.

a. Penrose's Gödelian Argument

Penrose has defended different versions of the Gödelian argument. In his earlier work, he defended a version of the argument that was relatively similar to Lucas's (although there were some minor differences (for example, in his argument, Penrose used Turing's theorem, which is closely related to Gödel's first incompleteness theorem)). Insofar as this version of the argument overlaps with Lucas's, this version faces many of the same objections as Lucas's argument. In his (1994) though, Penrose formulates a version of the argument that has some more significant differences from Lucas's version. Penrose regards this version "as the central (new) core argument against the computational modelling of mathematical understanding" offered in his (1994) and notes that some commentators seem to have completely missed the argument (Penrose 1996: 1.3).

Here is a summary of the new argument (this summary closely follows that given in Chalmers (1995: 3.2), as this is the clearest and most succinct formulation of the argument I know of): (1) suppose that "my reasoning powers are captured by some formal system *F*," and, given this assumption, "consider the class of statements I can know to be true." (2) Since I know that I am sound, *F* is sound, and so is *F*', which is simply *F* plus the assumption (made in (1)) that I am *F* (incidentally, a sound formal system is one in which only valid arguments can be proven). But then (3) "I know that G(*F*') is true, where this is the Gödel sentence of the system *F*'" (ibid). However, (4) Gödel's first incompleteness theorem shows that *F*' could not see that the Gödel sentence is true. Further, we can infer that (5) I am *F*' (since *F*' is merely *F* plus the assumption made in (1) that I am *F*), and we can also infer that I can see the truth of the Gödel sentence (and therefore given that we are *F*', *F*' can see the truth of the Gödel sentence). That is, (6) we have reached a contradiction (*F*' can both see the truth of the Gödel sentence and cannot see the truth of the Gödel sentence). Therefore, (7) our initial assumption must be false, that is, *F*, or any formal system whatsoever, cannot capture my reasoning powers.

Chalmers (1995: 3.6) thinks the "greatest vulnerability" with this version of the argument is step (2); specifically, he thinks the claim that we know that we are sound is problematic (he attempts to show that it leads to a contradiction (see Chalmers 1995: section 3)). Others aside from Chalmers also reject the claim that we know that we are sound, or else they reject the claim that we are sound to begin with (in which case we do not know that we are sound either since one cannot know a falsehood). For example, McCullough (1995: 3.2) claims that for Penrose's argument to succeed, two claims must be true: (1) "Human mathematical reasoning

is sound. That is, every statement that a competent human mathematician considers to be "unassailably true" actually is true," and (2) "The fact that human mathematical reasoning is sound is itself considered to be "unassailably true."" These claims seem implausible to McCullough (1995: 3.4) though, who remarks, "For people (such as me) who have a more relaxed attitude towards the possibility that their reasoning might be unsound, Penrose's argument doesn't carry as much weight." In short, McCullough (1995) thinks it is at least possible that mathematicians are unsound so we do not definitively know that mathematicians are sound. McDermott (1995) also questions this aspect (among others) of Penrose's argument. Looking at the way that mathematicians actually work, he (1995: 3.4) claims, "it is difficult to see how thinkers like these could even be remotely approximated by an inference system that chugs to a certifiably sound conclusion, prints it out, then turns itself off." For example, McDermott points out that in 1879 Kempe published a proof of the four-color theorem which was not disproved until 1890 by Heawood; that is, it appears there was an 11 year period where many competent mathematicians were unsound.

Penrose attempts to overcome such difficulties by distinguishing between individual, correctable mistakes that mathematicians sometimes make and things they know are "unassailably" true. He (1994: 157) claims "If [a] robot is...like a genuine mathematician, although it will still make mistakes from time to time, these mistakes will be correctable... according to its own internal criteria of "unassailable truth."" In other words, while mathematicians are fallible, they are still sound because their mistakes can be distinguished from things they know are unassailably true and can also be corrected (and any machine, if it is to mimic mathematical reasoning, must be the same way). The basic idea is that mathematicians can make mistakes and still be sound because only the unassailable truths are what matter; these truths are the output of a sound system, and we need not worry about the rest of the output of mathematicians. McDermott (1995) remains unconvinced; for example, he wonders what "unassailability" means in this context and thinks Penrose is far too vague on the subject. For more on these issues, including further responses to these objections from Penrose, see Penrose (1996).

b. Consciousness and Physics

One significant difference between Lucas's and Penrose's discussions of the Gödelian argument is that, as alluded to above, Penrose infers a number of further claims from the argument concerning consciousness and physics. Penrose thinks the Gödelian argument implies, for example, that consciousness must somehow arise from the quantum realm (specifically, from the quantum properties of "microtubules") and that we "will have no chance...[of understanding consciousness]... until we have a much more profound appreciation of the very

nature of time, space, and the laws that govern them" (Penrose 1994: 395). Many critics focus their attention on defeating Penrose's Gödelian argument, thinking that if it fails, we have little or no reason to endorse Penrose's claims about consciousness and physics. McDermott (1995: 2.2) remarks, "all the plausibility of Penrose's theory of "quantum consciousness" in Part II of the book depends on the Gödel argument being sound," so, if we can refute the Gödelian argument, we can easily reject the rest. Likewise, Chalmers (1995: 4.1) claims that the "reader who is not convinced by Penrose's Gödelian arguments is left with little reason to accept his claims that physics is non-computable and that quantum processes are essential to cognition..." While there is little doubt that Penrose's claims about consciousness and physics are largely motivated by the Gödelian argument, Penrose thinks that one might be led to such views in the absence of the Gödelian argument (for example, Penrose (1994) appeals to Libet's (1992) work in an effort to show that consciousness cannot be explained by classical physics). Some (such as Maudlin (1995)) doubt that there even is a link between the Gödelian argument and Penrose's claims about consciousness and physics; therefore, even if the Gödelian argument is sound, this might not imply that Penrose's views about consciousness and physics are true. Still others have offered objections that directly and specifically attack Penrose's claims about consciousness and physics, apart from his Gödelian argument; some of these objections are now briefly discussed.

Some have expressed doubts over whether quantum effects can influence neural processes. Klein (1995: 3.4) states "it will be difficult to find quantum effects in pre-firing neural activity" because the brain operates at too high of temperature and "is made of floppy material (the neural proteins can undergo an enormously large number of different types of vibration)." Furthermore, Penrose "discusses how microtubules can alter synaptic strengths...but nowhere is there any discussion of the nature of synaptic modulations that can be achieved quantum-mechanically but not classically" (Klein 1995: 3.6). Also, "the quantum nature of neural activity across the brain must be severely restricted, since Penrose concedes that neural firing is occurring classically" (Klein 1995: 3.6). In sum, at least given what we know at present, it is far from clear that events occurring at the quantum level can have any effect, or at least much of an effect, on events occurring at the neural level. Penrose (1994) hopes that the specific properties of microtubules can help overcome such issues.

As mentioned above, the Gödelian argument, if successful, would show that strong artificial intelligence is false, and of course Penrose thinks strong A.I. is false. However, Chalmers (1995: 4.2) argues that Penrose's skepticism about artificial intelligence is driven largely by the fact that "it is so hard to see how the mere enaction of a computation should give rise to an inner subjective life." But it isn't clear how locating the origin of consciousness in quantum processes that occur in microtubules is supposed to help: "Why should quantum processes in

microtubules give rise to consciousness, any more than computational processes should? Neither suggestion seems appreciably better off than the other" (ibid). According to Chalmers, Penrose has simply replaced one mystery with another. Chalmers (1995: 4.3) feels that "by the end of the book the "Missing Science of Consciousness" seems as far off as it ever was."

Baars (1995) has doubts that consciousness is even a problem in or for physics (of course, some philosophers have had similar doubts). Baars (1995: 1.3) writes,

> The...beings we see around us are the products of billions of years of biological evolution. We interact with them – with each other – at a level that is best described as psychological. All of our evidence regarding consciousness ...would seem to be exclusively psychobiological.

Furthermore, Baars cites much promising current scientific work on consciousness, points out that some of these current theories have not yet been disproven, that, relatively speaking, our attempt to explain consciousness scientifically is still in its infancy, and concludes that "Penrose's call for a scientific revolution seems premature at best" (Baars 1995: 2.3). Baars is also skeptical of the claim that the solution to the problem of consciousness will come from quantum mechanics specifically. He claims "there is no precedent for physicists deriving from [quantum mechanics] any macro-level phenomenon such as a chair or a flower...much less a nervous system with 100 billion neurons" (section 4.2) and remarks that it seems to be a leap of faith to think that quantum mechanics can unravel the mystery of consciousness.

4. Gödel's Own View

One interesting question that has not yet been addressed is: what did Gödel think his first incompleteness theorem implied about mechanism and the mind in general? Gödel, who discussed his views on this issue in his famous "Gibbs lecture" in 1951, stated,

> So the following disjunctive conclusion is inevitable: Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified . . . (Gödel 1995: 310).

That is, his result shows that either (i) the human mind is not a Turing machine or (ii) there are certain unsolvable mathematical problems. However, Lucas (1998: paragraph 1) goes even further and argues "it is clear that Gödel thought the second disjunct false," that is Gödel "was implicitly denying that any Turing machine could emulate the powers of the human mind." So,

perhaps the first thinker to endorse a version of the Lucas-Penrose argument was Gödel himself.

5. Other Anti-Mechanism Arguments

Finally, there are some alternative anti-mechanism arguments to Lucas-Penrose. Two are briefly mentioned. McCall (1999) has formulated an interesting argument. A Turing machine can only know what it can prove, and to a Turing machine, provability would be tantamount to truth. But Gödel's theorem seems to imply that truth is not always provability. The human mind can handle cases in which truth and provability diverge. A Turing machine, however, cannot. But then we cannot be Turing machines. A second alternative anti-mechanism argument is formulated in Cogburn and Megill (2010). They argue that, given certain central tenets of Intuitionism, the human mind cannot be a Turing machine.

6. References and Further Reading

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

Makes a number of objections to Lucas's argument; for example, the complexity of the human mind implies that we might be unable to formulate our own Gödel sentence.

Boyer, D. (1983). "J. R. Lucas, Kurt Godel, and Fred Astaire," *Philosophical Quarterly* 33:147-59.

Argues, among other things, that human output is finite and so can be simulated by a Turing machine.

Chalmers, D. J. (1996). "Minds, Machines, and Mathematics," *Psyche* 2:11-20.

Contra Penrose, we cannot know that we are sound.

Coder, D. (1969). "Gödel's Theorem and Mechanism," *Philosophy* 44:234-7.

Not everyone can understand Gödel, so Lucas's argument does not apply to everyone.

Cogburn, J. and Megill, J. (2010). "Are Turing machines Platonists? Inferentialism and the Philosophy of Mind," *Minds and Machines* 20(3): 423-40.

Intuitionism and Inferentialism entail the falsity of the Computational Theory of Mind.

Dennett, D.C. (1972). "Review of The Freedom of the Will," *The Journal of Philosophy* 69: 527-31.

Discusses Lucas's *The Freedom of the Will*, and specifically his Gödelian argument.

Feferman, S. (1996). "Penrose's Godelian argument," *Psyche* 2(7).

Points out some technical mistakes in Penrose's discussion of Gödel's first theorem. Penrose responds in his (1996).

Gödel, K. (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," *Monash. Math. Phys*. 38: 173-198.

Gödel's first incompleteness theorem.

Gödel, K. (1995). *Collected Works III* (ed. S. Feferman). New York: Oxford University Press.

Gödel discusses his first theorem and the human mind.

Dennett, D.C. and Hofstadter, D. R. (1981). *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Basic Books.

Contains Hofstadter's discussion of the Whiteley sentence.

Hutton, A. (1976). "This Gödel is Killing Me," *Philosophia* 3:135-44.

Probabilistic arguments that show that we can't know we are consistent.

Klein, S.A. "Is Quantum Mechanics Relevant to Understanding Consciousness," *Psyche* 2(3).
Questions Penrose's claims about consciousness arising from the quantum mechanical realm.

Lewis, D. (1969). "Lucas against Mechanism," *Philosophy* 44:231-3.
Lucas cannot produce all of "Lucas Arithmetic."

Libet, B. (1992). "The Neural Time-factor in Perception, volition and free will," *Review de Metaphysique et de Morale* 2:255-72.
Penrose appeals to Libet to show that classical physics cannot account for consciousness.

Lucas, J. R. (1961). "Minds, Machines and Gödel," *Philosophy* 36:112-127.
Lucas's first article on the Gödelian argument.

Lucas, J. R. (1968). "Satan Stultified: A Rrejoinder to Paul Benacerraf," *Monist* 52:145-58.
A response to Benacerraf's (1967).

Lucas, J. R. (1970a). "Mechanism: A Rejoinder," *Philosophy* 45:149-51.
Lucas's response to Coder (1969) and Lewis (1969).

Lucas, J. R. (1970b). *The Freedom of the Will*. Oxford: Oxford University Press.
Discusses and defends the Gödelian argument.

Lucas, J. R. (1976). "This Gödel is killing me: A rejoinder," *Philosophia* 6:145-8.
Lucas's reply to Hutton (1976).

Lucas, J. R. (1990). "Mind, machines and Gödel: A retrospect." A paper read to the Turing Conference at Brighton on April 6th.
Overview of the debate; Lucas considers numerous objections to his argument.

Lucas, J. R. (1996). "The Godelian Argument: Turn Over the Page." A paper read at a BSPS conference in Oxford.
Another overview of the debate.

Lucas, J. R. (1998). "The Implications of Gödel's Theorem." A paper read to the Sigma Club.
Another overview.

Nagel, E. and Newman J.R. (1958). *Gödel's Proof*. New York: New York University Press.
Short and clear introduction to Gödel's first incompleteness theorem.

Martin, J. and Engleman, K. (1990). "The Mind's I Has Two Eyes," *Philosophy* 510-16.
More on the Whiteley sentence.

Maudlin, T. (1996). "Between the Motion and the Act..." *Psyche* 2:40-51.
There is no connection between Penrose's Gödelian argument and his views on consciousness and physics.

McCall, S. (1999). "Can a Turing Machine Know that the Gödel Sentence is True?" *Journal of Philosophy* 96(10): 525-32.
An anti-mechanism argument.

McCullough, D. (1996). "Can Humans Escape Gödel?" *Psyche* 2:57-65.
Among other things, doubts that we know we are sound.

McDermott, D. (1996). "*Penrose is wrong," *Psyche* 2:66-82.
Criticizes Penrose on a number of issues, including the soundness of mathematicians.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.
Penrose's first book on the Gödelian argument and consciousness.

Penrose, R. (1994). *Shadows of the Mind*. Oxford: Oxford University Press.
Human reasoning cannot be captured by a formal system; consciousness arises from the quantum realm; we need a revolution in physics to fully understand consciousness.

Penrose, R. (1996). "Beyond the Doubting of a Shadow," *Psyche* 2(23).
Responds to various criticisms of his (1994).

Priest, G. (2003). "Inconsistent Arithmetic: Issues Technical and Philosophical," in *Trends in Logic: 50*

Years of Studia Logica (eds. V. F. Hendricks and J. Malinowski), Dordrecht: Kluwer Academic Publishers.

Discusses paraconistent logic.

Putnam, H. (1960). "Minds and Machines," *Dimensions of Mind. A Symposium* (ed. S. Hook). London: Collier-Macmillan.

Raises the consistency issue for Lucas.

Rogers, H. (1957). *Theory of Recursive Functions and Effective Computability* (mimeographed).

Early mention of the issue of consistency for Gödelian arguments.

Whitehead, A. N. and Russell, B. (1910, 1912, 1913). *Principia Mathematica*, 3 vols, Cambridge: Cambridge University Press.

An attempt to base mathematics on logic.

Wang, H. (1981). *Popular Lectures on Mathematical Logic*. Mineolam NY: Dover.

Textbook on formal logic.

Whiteley, C. (1962). "Minds, Machines and Gödel: A Reply to Mr. Lucas," *Philosophy* 37:61-62.

Humans are limited in ways similar to machines.

Wright, C. (1995). "Intuitionists are Not Turing Machines," *Philosophia Mathematica* 3(1):86-102.

An intuitionist who advances the Lucas-Penrose argument can overcome the worry over our consistency.

Author Information

Jason Megill
Email: jmegill@carroll.edu
Carroll College
U. S. A.