Intuitionists Are Not (Turing) Machines

CRISPIN WRIGHT*

I

In this paper I want to take up once again the widely discussed and disbelieved thesis of writers such as J. R. Lucas[1] and, more recently, Roger Penrose[2] who have argued that Gödel's incompleteness theorem for arithmetic provides as clear a demonstration as philosophy could reasonably hope for that our arithmetical capacities in particular, and hence the powers of the human intellect in general, cannot in principle be simulated by *machine*. I shall contend that of the two most influential objections to this thesis, one, first voiced, I think, by Paul Benacerraf,[3] is simply misconceived while the other, forcefully stated by Hilary Putnam,[4] may—at least by an intuitionistic proponent of Lucas/Penrose—be answered head-on. However, it seems to me that there remains cause for incredulity about the Lucas/Penrose line, to which I will come at the end.

II

Penrose expresses the basic idea like this:

> ... *whatever*... algorithm a mathematician might use to establish mathematical truth... there will always be mathematical propositions, such as the explicit Gödel proposition [for the formal system associated with that algorithm], that his algorithm cannot provide an answer for. If the workings of the mathematician's mind are entirely algorithmic, then the algorithm... that he actually uses to form his judgements is not capable of dealing with the [Gödelian]

* Department of Logic and Metaphysics, University of St Andrews, Fife KY16 9AL, Scotland.

[1] J. R. Lucas's classic paper is 'Minds, Machines and Gödel', *Philosophy* **36** (1963), 112-137.

[2] Roger Penrose *The Emperor's New Mind*, New York: Oxford University Press, 1989 (Vintage paperback edition 1990). See especially the section entitled 'The non-algorithmic nature of mathematical insight' in chapter 10, pp. 538–541. See also the sections, 'Gödel's theorem' and 'Mathematical insight' in chapter 4, pp. 138–146.

[3] Paul Benacerraf, 'God, the Devil and Gödel', *The Monist* **51** (1967), pp. 9–32.

[4] Hilary Putnam 'Minds and Machines' in volume 2 of his collected papers, *Mind, Language and Reality*, Cambridge: Cambridge University Press, 1975, pp. 362–385.

PHILOSOPHIA MATHEMATICA (3) Vol. 3 (1995), pp. 86–102.

> proposition... constructed from his personal algorithm. Nevertheless *we* can (in principle) see that [that proposition] is *true*. That would seem to provide *him* with a contradiction, since *he* ought to be able to see that also. Perhaps this indicates that the mathematician was *not* using an algorithm at all![5]

The central thought here is very simple: Gödel's theorem teaches us that the class of arithmetical truths ratifiable by the human mind does not coincide with those deliverable by any particular program for proof construction—any particular algorithm. Yet anything worth regarding as a *machine* has to be slave to such an algorithm. It follows that our arithmetical powers are not exhaustively mechanical.

A distinguished tradition of disbelief has grown up in response to this line of argument. Benacerraf's widely accepted objection referred to is that our ability to construct a Gödel sentence for a particular formal system, and hence to recognise the truth of that sentence, is of course hostage to our capacity intelligently to receive a specification of the formal system—algorithm—in question. The Lucas/Penrose argument, the objection goes, simply overlooks the possibility that human arithmetical capacity is indeed encoded in a particular formal system of which, however, we are unable to comprehend any formal specification suitable for an application of the Gödelian construction. It need not be questioned that we can produce a Gödel sentence for any arithmetical system which we can take in. But even granting that we can go on to recognise such a sentence to be true, what follows is merely a disjunction: *either* no such system encodes all human arithmetical capacity—the Lucas/Penrose thought—*or* any system which does has no axiomatic specification which human beings can comprehend. Lucas and Penrose simply overlook the latter disjunct.

Indeed, Gödel himself seems to have favoured essentially this conclusion, writing

> For if the human mind were equivalent to a finite machine,... there would exist *absolutely* unsolvable Diophantine problems of the type described above, where the epithet 'absolutely' means that they would be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof the human mind can conceive. So the following disjunctive conclusion is inevitable: ... *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.*[6]

But is this conservative—disjunctive—conclusion the most that is justified? It is never very clear what is meant by talk of the powers of the 'human

[5] *The Emperor's New Mind*, pp. 538–539

[6] From ms. p. 13 of Gödel's 1951 Josiah Willard Gibbs Lecture, delivered to the American Mathematical Society. See Kurt Gödel, *Collected Works*, volume III, edited by Solomon Feferman *et al.*, New York: Oxford University Press, forthcoming. For discussion of Gödel's view, see George Boolos's 'Introductory Note to #1951', *ibid.*

mind'. But in this context the unclarity is urgent. If we are talking about *actual* human minds, constrained as they are by the volume and powers of actual human brains, then it can hardly be doubted that the range of arithmetical sentences whose truth we are capable, by whatever means, of ratifying, is *finite*. So there are, of course, algorithms which capture all those sentences and more. Since Lucas and Penrose are presumably in no doubt about that, we had better conclude that the anti-mechanist thesis concerns the character of mathematical thought as reflected in what *in principle* we have the power to accomplish in mathematics. The comparison has to be not between *actual* human arithmetical capacities, even those of the most gifted and prodigious mathematicians, and *actual* theorem-proving machines, even the most powerful and sophisticated that technology will ever accomplish, but between what we can do *in principle* and what *in principle* can be accomplished by a machine, where a 'machine' is constrained only to the extent that its theorem-proving capacities correspond to the output of an effectively axiomatised system.

But what ought 'in principle' to mean for the purposes of such a comparison? I propose we should count any feat as within our compass in principle just in case some *finite* extension of powers we actually have would enable us to accomplish it *in practice*. Correspondingly for the powers in principle of a theorem-proving machine. So each first-order deductive consequence of the axioms which it encodes can in principle be proved by such a machine. Likewise, any arithmetical statement is ratifiable in principle by the human mind just in case, under some finite extension of capacities we actually possess, we could ratify it in practice.

Well plainly, if these are the terms of the intended comparison, then the mooted reservation about the Lucas/Penrose line lapses. There is no need for the second disjunct in the Gödelian statement. For any effectively axiomatised system of arithmetic is, in the relevant sense, *in principle surveyable* by human beings, who may consequently, in principle, carry out the Gödelian construction upon it. Even if in practice, limited as we actually are, our actual arithmetical capacities can be mirrored by a formal system whose specification we can neither comprehend nor, therefore, Gödelise, it remains that we have in principle the resources—a finite extension of our capacities would enable us—to do so. The qualification marked by the second disjunct in Gödel's statement is called for only if the words 'can conceive' concern human abilities in practice.

However, so to define the ground rules for the debate is likely, I imagine, to provoke the following complaint. Let it be, for the sake of argument, that, in the sense of 'in principle' stipulated, what the human mind can do in principle and what is in principle possible for a theorem-proving machine do indeed diverge. Still, the intention of the opponents of mechanism is surely to make good a claim which concerns not *semi-divine* creatures—creatures

whose powers in practice coincide with ours in principle—but actual living human beings. We were supposed to learn something about how to regard *ourselves*, limited as we are, and not about supermen who differ from us by having no finite limits of memory, rate of working, concentration, intellectual fluency, and so on.

Against this, the anti-mechanist can plausibly contend that it is just a mistake to think that if the claimed disanalogy in powers 'in principle' were sustained, we would still learn nothing about our actual limited selves. On the contrary, we would learn that, limited though we are, we work with a concept of arithmetical demonstration whose extension *allows no effective (recursive) enumeration*. By contrast, the extension of the concept of demonstration encoded in any formal system that might serve as the program for a theorem-proving machine *is* effectively enumerable. A satisfactory description of our actual arithmetical powers would thus have to acknowledge that they are informed by a concept of a sort by which, if our arithmetical thought were purely the implementation, as it were, of an in principle specifiable formal system, they would not be informed. Accordingly, if a disanalogy is pointed to at all by the Lucas/Penrose line, then it indeed is between human beings as we actually are and human beings as we would be if our mathematical thought were entirely algorithmic. The concept of demonstration which actually informs our construction and ratification of proofs would have been shown to have a feature which it would lack if our powers in those respects were purely mechanical.

III

That this is so is not questioned, however, by a second standard line of criticism of the Lucas/Penrose thesis. Rather, the second criticism rejects the supposition that we *are* able to 'see' the universally quantified undecidable sentence (U) to be true in the first place, that Gödel's construction does somehow provide or underpin anything worth regarding as an *informal demonstration* of U. Lucas and Penrose, the charge is, conveniently forget that any ground which emerges from Gödel's reasoning for accepting the truth of U is *entirely dependent* on the hypothesis of the consistency of the object system. And for that hypothesis Gödel's reasoning produces, of course, no grounds whatever. It may be a hypothesis in whose truth we are confident, and perhaps have some sort of good grounds for being so.[7]

[7] It might be maintained, for instance, that we have empirical-inductive reason to believe in the consistency of arithmetic. Compare Hartry Field on the consistency of set theory:

> ...a large part of the *reason* most of us believe that modern set theory is consistent is the thought that if it weren't consistent someone would probably have discovered an inconsistency by now. (*Realism, Mathematics & Modality*, Oxford: Blackwell, 1989, p. 232.)

Whatever grounds those may be then transfer into grounds for believing the truth of U. But that's a far cry from saying that something worth regarding as a mathematical *demonstration* of U is in the offing. Thus Hilary Putnam writes:

> Let T be the Turing machine which [purportedly] 'represents' me in the sense that T can prove just the mathematical statements I can prove. Then the argument... is that by using Gödel's technique I can discover a proposition that T cannot prove, and moreover I can prove this proposition. This refutes the assumption that T 'represents' me, hence I am not a Turing machine. The fallacy is a misapplication of Gödel's theorem, pure and simple. Given an arbitrary machine T, all I can do is find a proposition U such that I can prove:
>
> (3) If T is consistent, U is true,
>
> where U is undecidable by T if T is in fact consistent. However, T can perfectly well prove (3) too! And the statement U, which T *cannot* prove (assuming consistency), I cannot prove either (unless I can prove that T is consistent, which is unlikely if T is very complicated.)[8]

On this view, no disanalogy supportive of the Lucas/Penrose line emerges from consideration of Gödel's theorem. Sure, for any recursively axiomatised, intuitively acceptable arithmetical system, a human mathematician can in principle 'demonstrate' the relevant U by means which cannot be mimicked in that system (if consistent)—*provided* his 'demonstration' is allowed to help itself to the assumption of the consistency of the relevant system. But the intended disanalogy was to be that the means which the human mind can in principle deploy for, as Penrose put it, *establishing mathematical truth*, transcends anything available to (anything worth regarding as) a machine. So the mooted kind of *modus ponens* has to qualify as *establishing the truth* of U, and hence must draw only on premises which may themselves be regarded as established. Where is the argument for thinking of the consistency of arithmetic as such an established truth?

The question is not rhetorical. One can envisage a direct reply. It might responsibly be argued, for instance, either that the truth, hence consistency of standard axioms for first-order arithmetic should be regarded as primitively a priori certain, or that their consistency is in effect established by

Likewise:

> ... much of our knowledge of possibility is to some extent inductive. For instance, our knowledge that [von Neumann–Bernays–Gödel set theory is consistent] seems to be based in part on the fact that we have been unable to find any inconsistency in [that system]. (*Ibid.*, p. 88.)

I am sceptical about this general line. But it would be irrelevant to our present purpose to pursue it here.

[8] 'Minds and Machines', p. 366.

Gentzen's famous proof.[9] But it would be much better from the point of view of Lucas and Penrose if a form of demonstration of the Gödel sentence could be disclosed which, while applicable to any intuitively acceptable system of arithmetic, involved no reliance on the supposition of its consistency as a *premise*. Just that is what is promised by the specifically intuitionistic form of demonstration to which I now turn.[10]

IV

The key ingredients in the demonstration are the intuitionists' official account of negation and the reflection that Gödel provides the resources for finding a contradiction in an arithmetical system, S, if there is a counterexample to the appropriate U. The claim is that these considerations, suit-

[9] As Gentzen himself thought:

> ...I am inclined to believe that in terms of the fundamental distinction between disputable and indisputable methods of proof, the *proof of the finiteness* of the reduction procedure

—the part of his proof that uses transfinite induction—

> can still be considered indisputable, so that the consistency proof represents a *real vindication* of the disputable parts of elementary number theory.

(from chapter 4, 'The consistency of elementary number theory', of *The Collected Papers of Gerhard Gentzen*, edited by M. Szabo, Amsterdam: North Holland, 1969, p. 197)

The issue is certainly subtle and deserves detailed discussion. But it is hard not to feel sceptical about Gentzen's view. A derivation only counts as a demonstration in the sense in which we are interested if its premises and the rules of proof which it exploits are, in a sense I shall not here attempt to define, antecedently epistemically more basic—more secure—than the purported conclusion. Now we know from Gödel's second incompleteness theorem that no consistent arithmetical system can contain the resources for a proof of its own consistency. The question is therefore whether the additional resources deployed in any such proof—essentially the use made of transfinite induction up to ε_0 in the case of the Gentzen proof—may rightly be regarded as epistemically more basic than the proof theory, and especially the full-fledged induction schema, of first-order number theory itself. Well, to accept the Gentzen proof is to be persuaded of a mapping between the proofs constructible in elementary number theory and the series of ordinals up to ε_0. And to understand the structure of the ordinals up to ε_0 is to grasp a concept which embeds and *builds on* an ordinary understanding of the series of natural numbers. So to treat the Gentzen proof as establishing consistency is implicitly to exclude any antecedent doubt about the *coherence* of the concept of natural number. Any permissible antecedent doubt about the consistency of first-order number theory would have, therefore, to concern whether the (fully coherent) concept of natural number is faithfully reflected by the standard first-order axioms—specifically, by all admissible instances of the induction schema. Such a doubt would have to concern whether, even when restricted to first-order arithmetical vocabulary, 'tricksy' predicates might not somehow be formulated whose use in inductions could lead to contradiction. Could someone reasonably worry about that who was confident in the consistency of Gentzen's methods? How might such a doubt be further elaborated?

[10] Prescinding from the intuitionistic demonstration which we are about to consider, I suspect the real reason why so many of us tend to regard U as informally demonstrated by Gödel's construction is actually rather unflattering, *viz.* we succumb to a simple conflation, confusing the discovery of a *commitment* with the discovery of a *truth*. The proof of Putnam's conditional (3) for an intuitively correct S is a deeply impressive result which teaches us, on pain of accepting the inconsistency of our arithmetical thought, that

ably deployed, provide intuitionistic grounds for the affirmation of U which involve no inference, implicit or explicit, from the presumed consistency of S. Too good to be true?

In order to set matters up, we need to review a little of the general detail of Gödel's construction. As is familiar, the groundwork for his result is the assignment to each primitive symbol in the vocabulary of S, and thereby to each well-formed formula and sequence of such formulae, of a number—its *Gödel-number*. This assignment is two-way effective; *i.e.*, given an arbitrary formula, or sequence of formulae of S, we can in principle effectively compute its Gödel-number; and given an arbitrary number, we can effectively determine whether it is the Gödel-number either of a formula or a sequence of formulae of S and if so, to which formula or sequence of formulae it belongs. Crucially, as Gödel shows, such an assignment may be done in such a way that a certain primitive recursive relation, Pxy, holds between the numbers x and y just in case x is the Gödel-number of a sequence of formulae which constitutes a formal proof in S of the formula whose Gödel-number is y. That is, for all x and y,

(Lemma) Pxy *iff* y *is the Gödel-number of a wff,* B, *of* S *and* x *is the Gödel-number of a sequence of wffs of* S *which constitute a proof in* S *of* B.

The proof now proceeds by showing how the operation,

> substitution for each occurrence of the free variable with Gödel-number m in the formula whose Gödel-number is k by the numeral for n,

may be arithmetically represented; that is, how an effective function in three arguments, $Sub\langle k, m, n\rangle$, may be arithmetically defined in such a way that, when k is the Gödel-number of a formula containing free occurrences of the variable with Gödel-number m, the value of $Sub\langle k, m, n\rangle$ will precisely be the Gödel-number of the formula that results from substituting the numeral

we are *committed* to regarding each instance of U as computationally verifiable. Since we want to believe that our arithmetical thought is consistent—since, indeed, it is doubtful if agnosticism on the matter could be a practical option—there can therefore be no sitting on the fence as far as U is concerned. But it is quite another thing to view this commitment to U as something we incur on the basis of a *demonstration* of its truth. We are, as a rough parallel, similarly committed by everything we ordinarily think and do to the existence of the material world; no agnosticism on the point is practical. It would be very much easier than it is to dispose of material-world scepticism if this commitment could immediately be taken as the reflection of a cognitive achievement.

If this somewhat deflating account is correct, we are obliged to conclude that no informal demonstration of the undecidable sentence attends Gödel's proof, and nothing takes place worthy of dignification as 'recognition' of its truth. We are merely brought to see that our standing commitment to the consistency of our arithmetical thought embraces a plethora of unsuspected, specifically arithmetical commitments, each associated with a Gödelian undecidable sentence. That is enormously interesting. But it provides no support whatever for Lucas and Penrose.

for n for each free occurrence of that variable. We then consider the open sentence

$$(U^*) \qquad (\forall x) \sim (Px, Sub\langle y, m, y\rangle),$$

where m is the Gödel-number assigned to the variable 'y'. This will have a Gödel-number—say, g. So '$Sub\langle g, m, g\rangle$' will denote the Gödel-number of the formula that results from substitution for each occurrence of the free variable with Gödel-number m in the formula whose Gödel-number is g, *viz.* (U^*), by the numeral for g; *i.e.*, '$Sub\langle g, m, g\rangle$' will denote the Gödel-number of the formula

$$(U) \qquad (\forall x) \sim (Px, Sub\langle g, m, g\rangle),$$

—the undecidable sentence itself.

Now recall the standard informal explanation of intuitionistic negation:

Negation *The negation of* A *is demonstrated by any construction which demonstrates that a contradiction could be demonstrated if we had a demonstration of* A.[11]

The focus of what follows will be on the consequences of the supposition that an intuitively correct arithmetical system S contains a proof of 'Pkg' for some particular choice of 'k'. But to get any mileage, we now need to specify some assumptions about S. S is to be any formal arithmetic with a standard intuitionist logic which is strong enough to be within the scope of Gödel's theorems and such that:

(1) S is *intuitionistically endorsed*, *i.e.* all the methods and assumptions of S are intuitionistically acceptable, so that every proof in S corresponds to an intuitionistic demonstration (I-demonstration). Note: this is *not* to assume the consistency of S; what is entailed is merely that if S is inconsistent, then so are the principles and assumptions incorporated in the informal notion of I-demonstrability. And

(2) S is *computationally adequate*—all arithmetical computations can be done in S. In particular, for each primitive recursive 'Axy', and arbitrary 'k' and 'n', either 'Akn' or '$\sim Akn$' is computationally decidable in S. Again, reflect that S can, of course, be inconsistent while having this feature.

[11] Thus Heyting in his *Intuitionism: An Introduction* (Amsterdam: North Holland, 3rd edition, 1971):

> $\neg P$ can be asserted if and only if we possess a construction which from the supposition that a construction [proving] P were carried out, leads to a contradiction. (p. 102)

Compare Dummett in his *Elements of Intuitionism* (Oxford: Clarendon Press, 1977):

> A proof of $\neg A$ is usually characterised as a construction of which we can recognise that, applied to any proof of A, it will yield a proof of a contradiction. (p. 13)

(1) and (2) will be true of any formal arithmetic in which the intuitionist mathematician will be interested.

Now suppose that

(i) 'Pkg' is computationally verifiable.

Then (ii) 'Pkg' is provable in S—by (2).

But (iii) U is provable in S—from (i) by Gödel's *Lemma* above.

So (iv) '$\sim Pkg$' is provable in S—since U is provable in S and S has a standard intuitionist logic.

So (v) '$Pkg \,\&\, \sim Pkg$' is provable in S—by (ii) and (iv).

It has thus been shown that, had we a computational verification of 'Pkg', we could accomplish a proof in S, and hence, by (1), an I-demonstration, of a contradiction. But any I-demonstration of 'Pkg' will be, in view of the primitive recursiveness of the predicate, a demonstration of its computational verifiability, *i.e.*, of (i). By the foregoing, it will therefore constitute a demonstration of (v), so by (1) a demonstration that a contradiction is demonstrable. So, in the light of *Negation*, (i)–(v) constitutes an I-demonstration of '$\sim Pkg$'.

If that is right, it would seem that nothing obstructs advancing to a demonstration of U itself—indeed that the further step is merely trivial. Intuitionistically a demonstration of arithmetical '$\forall x(Ax)$' is any construction which we can recognise may be used, for an arbitrary natural number k, to accomplish a demonstration of 'Ak'.[12] (Call this principle *Generality*.) Well, the reasoning from (i) to (v) will evidently go through, if at all, then for an arbitrary choice of 'k'. So it ought to be acknowledged to constitute, by *Generality*, an I-demonstration of U. *QED*.

It appears, then, that the reasoning from (i) to (v) constitutes, on the assumptions made, the basis of an intuitionistically cogent inference to U. That will constitute no advance on what we have already, of course, if those assumptions somehow smuggle in a presupposition of consistency, either of the system S or more generally, of the informal collection of principles and axioms which are sanctioned by the intuitionistic notion of demonstration. But, as noted, this seems not to be so.

Why present this as a distinctively *intuitionistic* demonstration—what happens if we try to transpose the reasoning to provide for a *classical* demonstration of U? The crux is the part played by the intuitionistic account of negation—that which licenses the claim that you have demonstrated '$\sim A$' as soon as you have shown how, given any demonstration of 'A', a demonstration could then be given that a contradiction was demonstrable. Naturally, this account will not be acceptable to the classicist in all cases; from a classical point of view, truth is one thing and demonstrability quite another and there is in principle no reason why a *true* sentence

[12] *Cf.* Heyting, *op. cit.*, p. 106, and Dummett, *Elements of Intuitionism*, p. 12.

should not merely be absolutely unprovable but even such that if we could prove it, we could derive inconsistencies. However, might the intuitionistic account be classically acceptable in the relevant case, where A is 'Pkg' and therefore effectively decidable?

Well, the difference the effective decidability of 'Pkg' makes is that either it or its negation is thereby guaranteed to be classically demonstrable. A proof that either cannot be demonstrated is therefore a proof that the other can be. Now, we have in (i) to (v) a classically acceptable proof that, had we a demonstration of 'Pkg', we could demonstrate the obtainability of a contradiction. *If* that amounted to a proof that no demonstration of 'Pkg' can be given, then a classical demonstration that each '$\sim Pkg$' is demonstrable, hence one of U, would be in prospect. But the problem is clear: no such conclusion follows unless we assume that it is classically impossible to demonstrate the classical obtainability of a contradiction. And that is tantamount to the forbidden assumption that the methods and principles incorporated in the classical notion of demonstration are consistent—a stronger assumption, indeed, than the antecedent of Putnam's conditional (3).

There appears, then, to be an asymmetry between the intuitionistic and classical cases. Intuitionistic demonstrations with no classical analogue are always interesting—more than interesting when the conclusion is arithmetical. Can it *really* be that we have one here?

V

Experience has taught that this reasoning is liable to provoke more than its share of misconceived objections. But there are two doubts which are worth an immediate airing. The first concerns the role of *Negation* and focuses on assumption (1): that the system S is *intuitionistically endorsed*—that its postulates and deductive apparatus involve nothing that is not intuitionistically acceptable. As noted, if this assumption is to fall short of an assumption of consistency, then it can involve no claim about what intuitionists ideally *ought* to accept—since they certainly ought not to accept inconsistent systems—but must be restricted to the effect that the postulates and inference-rules of S all be ones which intuitionists have *historically accepted* or which are sanctioned by their historically given explanations. Correspondingly, therefore, the notion of an I-demonstration featuring in the conclusion of the reasoning from (i) to (v) above—*viz.* that had we a computational verification of 'Pkg', we could accomplish a proof in S, and hence, by (1), an I-demonstration, of a contradiction—must be understood to involve no more than conformity with such historically given explanations. Accordingly, if it is to subserve the next step, then the principle of *Negation*, it may seem, has to be interpreted somewhat along the following lines:

Negation* *If a construction is given which provides, by standards historically accepted by intuitionists, a demonstration that a contradiction could be demonstrated (by those same standards) if we had such a demonstration of* A, *then that construction constitutes a demonstration of the negation of* A.

But would intuitionists accept *this* principle? Ought they not to be alive to the possibility that the availability—by conventional intuitionist standards —of a demonstration of a contradiction on the hypothesis of such a demonstration of A, should be taken as reflecting not on the status of A but on the standards that sanction the putative demonstration? That is not an option if it is assumed that those standards can be met only by genuinely cogent demonstrations. But that is exactly what we cannot assume if we are to avoid presupposition of (anything stronger than) the consistency of S. In short: if S's being intuitionistically endorsed is interpreted in the weak sense explicitly stated, then either *Negation* is not acceptable or else it embeds a forbidden presupposition of the consistency of the informal intuitionistic notion of demonstration, and hence of S.[13]

The immediate reply to this is that it depends upon an equivocation. Specifically, it depends on interpreting the last occurrence of 'demonstration' in the reformulated *Negation** in an absolute, factive sense—the usual sense, whereby to have given a demonstration of a proposition is to have advanced rationally compelling grounds for accepting what is in fact a truth—while interpreting the others in the historic sense correlative to intuitionistic endorsement. Under this hybrid interpretation, it is indeed a possibility—should the principles traditionally accepted by intuitionists be in fact incoherent—that a construction complying with the antecedent of *Negation** should fall short of the demonstration advertised by its consequent. But first, it is by no means evident that the recognition of this possibility constitutes sufficient grounds to reject the hybrid principle—a construction's satisfaction of the antecedent may still be taken to make a sufficient, though defeasible case for regarding A as refuted; and second, the reflection is in any case quite beside the point since if a reinterpretation of the concept of demonstration is called for at all, it is called for *uniformly*. And *Negation** is uncontentious once *all* its references to demonstration are understood historically—it merely articulates what intuitionists have as a matter of fact been prepared to accept as demonstrations of negated statements.

However the objection may be pressed further. The argument of section IV was supposed to marshal considerations which, for an intuitionist, should be regarded as establishing the truth of the Gödel sentence for S. But once the concept of I-demonstration is uniformly understood in the historical

[13] This reservation has been expressed by Charles Parsons.

sense, then is not the upshot going to be at best an *historically* acceptable demonstration of the Gödel sentence for S? And if one is to take that as licensing the claim to have *established*—to have given rationally compelling reason for accepting—the Gödel sentence, must one not in effect presuppose the soundness, and hence consistency of the historically accepted methods—and thereby once again the soundness, and consistency, of S itself?

The answer is that one must, but that presupposition of soundness, and hence consistency, in *that* way ought not to be reckoned germane. It has to be granted that in treating *any* construction as a genuine (absolute) demonstration, a thinker implicitly presupposes the soundness of the premises and principles which it utilises. But that had better not be to admit that the soundness of that apparatus features as an *additional* premise—as if all we ever really succeed in demonstrating are conditional statements whose antecedents hypothesise the soundness of the methods deployed. The brief was to accomplish a demonstration of the Gödel sentence for S which—in contrast, for instance, with a *modus ponens* on Putnam's conditional (3)—makes no undischarged use of the assumption of S's consistency (or something stronger). If the best that a critic of the argument of section IV can do is point to a form of reliance upon such an assumption which is actually ubiquitous—is of the very essence of the market in proofs among rational thinkers who recognise their own fallibility—we should reckon that the brief has been fulfilled.

The second doubt arises as soon as we reflect that the intuitionist ought not to be satisfied with *Generality* as formulated above. For that formulation puts no specific controls on demonstrations of 'Ak'. What is required is, rather, something along the lines:

A demonstration of arithmetical '∀x(Ax)' *is any construction which we can recognise may be used, for an arbitrary numeral* 'k', *to accomplish a constructive demonstration of* 'Ak'.

Now, what is a constructive demonstration in the case where 'Ak' is decidable by computation if not (a guarantee of the possibility of) the appropriate computation? Since '$\sim Pkg$' is such a case, the reasoning in the intuitionistic demonstration, while admittedly applicable to an arbitrary numeral 'k', ought to be reckoned as providing a basis for the universal introduction step to U only if it ensures that each '$\sim Pkg$' is *verifiable by computation*. Does it?

Well, suppose the wider notion of I-demonstrability *is*, alas, inconsistent, although the fragment consisting of arithmetical computation is consistent. Let the system S be inconsistent, and 'Pkg' actually computationally verifiable. Then the informal demonstration of '$\sim Pkg$' still goes through, but now provides no assurance of computational verifiability. Hence, to treat it as providing such an assurance is implicitly to assume that our supposition

does not obtain—*i.e.*, that I-demonstrability is consistent. Either, then, the universal-introduction step is unjustified, since we have no assurance that each '$\sim Pkg$' is constructively verifiable; or, in taking it that we have such an assurance, we implicitly assume the consistency of I-demonstrability—and the forbidden assumption is implicitly there after all.

This is a good objection as far as it goes. It rests, however, on the proposal that, where A is effectively decidable by computation, a constructive proof of A must consist in, or anyway establish the possibility of, the appropriate computation. And other proposals are possible. Not all computationally decidable statements are atomic. So one possible proposal would be to grant that, in the atomic case, a constructive demonstration must consist in (proving the possibility of) the appropriate computation; but to allow, for certain non-atomic forms of computationally decidable statement, that certain kinds of proof besides the relevant kind of computation may count as constructive. In the crucial case, in particular, of the negations of atomic statements, it might be held to be sufficient for constructivity that (one prove that) *effective* means exist for arriving at a contradiction should a computational verification be provided of the statement negated. Since the decoding of 'Pkg' and consequent location of a proof in S of U, is an effective procedure, this more relaxed account would sanction the constructivity of the mode of demonstration outlined for each '$\sim Pkg$', and would—apparently without the need for the forbidden assumption—thus reinstate the intuitionistic demonstration of U under the aegis of the tightened version of *Generality*.

The status of the intuitionistic demonstration turns, then, on the vexed question of the proper interpretation of the notion of constructive proof—specifically on whether the intuitionist ought or ought not to allow that the more relaxed notion of constructivity outlined is what is relevant to '$\sim Pkg$'.

I shall not attempt to pursue the matter to a conclusion here. But by way of an opening salvo, here are three considerations whereby a supporter of the intuitionistic demonstration might at least begin to try to motivate the more relaxed notion of constructivity which he requires.

(i) An opponent of the more relaxed account can hardly argue that it misclassifies demonstrations as constructive which are actually no such thing. For, as the reader will speedily verify, the two accounts are coextensive unless I-demonstrability is inconsistent. So an argument for the extensional inadequacy of the relaxed account will need the inconsistency of intuitionistic mathematics as a premise.

(ii) Don't we need the more relaxed notion to make sense of the very problem case described? For that is a case where there is, by hypothesis, an I-demonstration—ergo a *constructive* demonstration—of U which

can be given in S; and whatever else that is, it is required—by tightened *Generality*—to involve a construction which enables us to provide a *constructive* demonstration of '$\sim Pkg$', for arbitrary k. Yet the problem case is supposed to be one where no computational verification of '$\sim Pkg$' is possible. So either the problem case makes no intuitionistic sense, or the notion of constructive demonstration cannot be restricted in the way required to make a problem in the first place!

(iii) It will not in general be sensible to rate proofs of truth-functional compounds of recursively decidable atomic statements as constructive only if they are constituted by the execution of the same class of recursive procedures. Such a proposal would restrict constructive proofs of, for instance, conditionals composed out of such statements to computational verifications of the consequent or falsifications of the antecedent—in no other case could such a conditional be regarded as constructively proved. And that would be to take all the conditionality out of the conditional, as it were—to abrogate the right to conditional claims in cases where one had no proof or disproof of either antecedent or consequent. The right account should allow such a conditional to have been constructively proved just in case we have shown how, given any computational verification of the antecedent, we could effectively get a computational verification of the consequent. If such a relaxation of the notion of constructive proof is appropriate for the conditional, why not for negation too?

VI

Provided it is intuitionistic number-theory with which we are concerned, then, there is at least a case to answer for the point that constitutes the nerve of the Lucas/Penrose argument: Gödel's theorem *can* be harnessed to an argument that there is no effective and complete enumeration of the extension of the concept of demonstration which informs (intuitionistic) human mathematical practice. So that concept has a feature which, no matter what Turing machine we consider, is lacked by the concept of demonstration instantiated by the output of that machine. Given any effective enumeration (recursive axiomatisation) of arithmetical truths, the claim is, we have a method—contained in the technique for constructing the Gödel sentence and then applying the general form of the intuitionistic demonstration to it—for generating a new demonstration going beyond what can be accomplished, even in principle, by derivations from the axioms in question. In brief, the structure of the output-in-principle of the human (intuitionist) mathematician and that of the output-in-principle of any Turing machine are different.

But can the non-mechanical character of (intuitionistic) human mathematical thought really be carried by this claim? So much artillery has been directed against the anti-mechanist argument *to* the claimed disanalogy

that neglect has resulted of the way the argument is supposed to proceed *from* it. But an (intuitionistic) advocate of the Lucas/Penrose line is by no means home.

First, a small but, in a context in which there has been much confusion about the role of suppositions of consistency, important clarificatory qualification: the indicated disanalogy can be made out *only* if we take it that human arithmetical thought is consistent. Otherwise there is *of course* a Turing machine which generates all and only the arithmetical sentences of which we can in principle construct what, by our standards, rank as (intuitionistic) demonstrations. Even granting that the form of a demonstration of U has been disclosed in which the consistency of the object system does not feature as a premise, the claim to have shown thereby that, in general, the class of in principle humanly demonstrable arithmetical sentences is not effectively enumerable, will still depend on the assumption of the consistency of that system and, indeed, of any intuitionistically acceptable arithmetic which strengthens it. The most that is in prospect, in other words, is still a *disjunctive* conclusion. Only the disjunction is not the Gödelian disjunction cited earlier. That disjunction featured as its second disjunct the proposition that the Turing machine which in fact encodes human arithmetical capacity is one whose formal specification no human being can comprehend. By contrast, the disjunction in prospect as the proper conclusion of the Lucas/Penrose line of thought will replace that disjunct with the (depressing) proposition that arithmetical demonstrability by intuitionistically acceptable means is an inconsistent notion.

This *is* a small qualification. Clearly it is not a terribly damaging concession for an intuitionist proponent of Lucas and Penrose to have to make if the conclusion has to be not that:

Human arithmetical thought is non-mechanical,

but that:

Human arithmetical thought, if not inconsistent, is non-mechanical.

The latter is still a claim of considerable philosophical interest.

There is, however, a further question about the attainability of even this qualified conclusion on which our discussion has so far not impinged, and which seems to me very moot. What we are assuming to be in prospect is a disanalogy, on the assumption of the consistency of the principles and methods which are intuitively acceptable to intuitionistic mathematicians, between the concept of demonstrability defined by those principles and methods and any concept of demonstrability which governs the workings of a *Turing* machine. Since, as was stressed earlier, the feasible arithmetical output, so to speak, of even the most prodigious human mathematician can no doubt be matched and indeed surpassed by a suitable Turing machine, making the disanalogy good will require reflection on the *intensions* of the

relevant concepts of demonstrability. A sufficiently explicit characterisation will therefore be needed of the human notion, so to speak, to make it clear how, for any particular arithmetical Turing machine—still assuming consistency—an arithmetical demonstration lying beyond its scope might in principle effectively be found. Well, suppose that accomplished in the preceding section. Then the basis of the Lucas/Penrose thesis will have consisted in nothing other than the provision of an *effective* procedure for finding, for any particular recursive axiomatisation of arithmetical demonstrations, an intuitively acceptable arithmetical demonstration not included within it.

Rather than striking a blow against mechanistic conceptions of the human intellect, there will therefore be an immediate question whether this whole trend of thought cannot at most disclose the inadequacy of the idea of a *Turing machine* as a stalking horse for mechanism. What an argument against mechanism ought to show is that, for the area of human thought where the mechanist thesis is contested, *insight imagination* and *creativity* have a role to play which cannot be simulated by mechanical model—which cannot be reduced to the implementation of any set of effective instructions. The great difficulty, always, is to render such ideas sufficiently precise to make them debatable, to make it clear what a defender or an opponent has to establish. But surely it is dubious whether the debate as envisaged has succeeded in doing that. Whatever is shown by an argument which establishes that, for any particular recursive axiomatisation of arithmetical proofs, there exists—if human arithmetical thought is consistent—an *effective* procedure for constructing a demonstration of an arithmetical sentence not included in the list, it is *not* that human thought is essentially creative, gifted with a spark which transcends the merely mechanical implementation of any instructions which can be laid down in advance. Any consistent, recursively axiomatised system of arithmetic may be so specified that its Gödelisation *is* an effective procedure. And to the sentence which results from that procedure may then, as it were mindlessly, be applied the form of demonstration outlined in the preceding section or whatever is the general form of the informal demonstration we are assuming has been provided.

True, the sentences which result from indefinite iteration of this procedure on an intuitively acceptable base arithmetic—say standard first-order Peano arithmetic—will not be recursively axiomatisable—will not coincide with the output of any particular Turing machine. But a proponent of Lucas and Penrose needs to say something to disarm the obvious response, *viz.* that a device or organism may therefore *still* be lacking that which opponents of mechanism wish to claim for the human mind even though there is no recursive enumeration of specifications of all the tasks which it is able to perform. Let it be that the Benacerraf objection is rightly discounted for the reason I indicated, and that the intuitionistic demonstration of U is acceptable. Then provided—however unlikely it might be in the case of

these actual protagonists!—Lucas and Penrose go intuitionist, they can win by the rules of debate they set themselves to follow The remaining doubt concerns the sufficiency of those rules to ensure that their victory means what it was intended to.[14]

ABSTRACT. Lucas and Penrose have contended that, by displaying how any characterisation of arithmetical proof programmable into a machine allows of diagonalisation, generating a humanly recognisable proof which eludes that characterisation, Gödel's incompleteness theorem rules out any purely mechanical model of the human intellect. The main criticisms of this argument have been that the proof generated by diagonalisation (i) will not be humanly recognisable unless humans can grasp the specification of the object-system (Benacerraf); and (ii) counts as a proof only on the (unproven) hypothesis that the object system is consistent (Putnam). The present paper argues that criticism (ii) may be met head-on by an intuitionistic proponent of the anti-mechanist argument; and that criticism (i) is simply mistaken. However the paper concludes by questioning the sufficiency of the situation for an interesting anti-mechanist conclusion.

[14] Thanks to Bob Hale, Stewart Shapiro and Neil Tennant for criticisms of an earlier draft, and to the participants at the Conference on the philosophy of Michael Dummett held at Mussomeli, Sicily, in September 1992 at which a version of the principal argument was presented.