Macalester Journal of Philosophy

Volume 7
Issue 1 *Spring, 1997*

Article 6

11-22-2010

A Discussion of Some Consequences of Gödel's Incompleteness Theorems

Sebastian Lange
Macalester College

Sebastian Lange

"A Discussion of Some Consequences of Gödel's Incompleteness Theorems"

Introduction

In the beginning of this century the famous Hilbert program was postulated--a program to revise the standards and practices of mathematics, an attempt to base mathematics on formal foundations. In essence, mathematicians hoped to find a system which would provably capture all mathematical truths. It was these hopes that brought Whitehead and Russell to create their seminal contribution to mathematics--the *Principia Mathematica.*

Yet, Kurt Gödel's proof upset all hopes of reaching this tempting formalistic paradise of an all truth capturing, internally provably consistent system. What can be thought of as the most simple and rudimentary part of mathematics, arithmetic, turned out to be incomplete, which means that not all truth of the system can be proven within the system itself. In fact, Gödel showed that any formal system strong enough to express arithmetic is necessarily incomplete, and that whenever we add those formerly unprovable statements (which were instantiations of unprovable truth) as a proper axiom to our system, we could nevertheless reapply Gödel's proof method to the new system to show it incomplete. In this essay I will try to look at some consequences of the proof in the fields of mathematics and artificial intelligence. I will claim that the proof has some great importance for the practice, as well as foundations, of mathematics. However, as regards artificial intelligence, I will suggest that Gödel's proof does not present a serious challenge.

What does Gödel's proof look like? Gödel's proof is based on the ancient Liar Paradox stemming from the expression, "This sentence is false." If this sentence were true, then it would express a truth, but what it expresses is that it is false. Thus, when we assume it is true, the sentence says about itself that it is false. On the other hand, if we assume that this sentence is false, then what it says must be false. But it says that it itself is false, whose negation is that it is true. Thus, when we assume it is false, it says about itself that it is true. Thus, if the sentence is true, it is false; and when it is false, it is true. We arrive at an impasse. Gödel showed that in any formal system strong enough to express arithmetic there exists an arithmetical statement which is interpreted as "I am unprovable." If this statement could be proven to be true, then we would prove a statement which is stating of itself that it is unprovable. Thus, we would arrive at a contradiction. However, if we assumed that we could prove this statement to be false, then we proved that what it says must be false. But it says that it is not provable; thus, we would prove that the above statement is provable. But we have just shown that the above statement is not provable, since it leads to a contradiction. Thus, we have no way of proving the statement either way. Yet, it really is true, since we have no way of proving it. This result leads to the two incompleteness theorems of Gödel. First, as we could gather from the summary of the proof above, there exists at least one statement in a formal system of arithmetic which is true, but cannot be proven within this system.

Secondly, Gödel used this result to show that we cannot prove the consistency of a formal system of arithmetic within the system itself.

There is what Hofstadter called a certain critical mass for a formal system (Hofstadter, 450). Once the system has reached an internal richness that fulfills the following conditions, the system is prey to the Gödel proof method showing it incomplete:

> (1) the system should be rich enough so that all desired statements about numbers, whether true or false, can be expressed in it...
>
> (2) That all general recursive relations should be represented by the formulas of the system ...
>
> (3) That the axioms and typographical patterns defined by its rules be recognizable by some terminating decision procedure (*Ibid.*).

Thus it seems that it is impossible to have it both ways: a system that is powerful enough to be mathematically interesting, i.e., capable of expressing all of arithmetic, and a system that will provably yield us all truths of the interesting domain. Let us now look closer at the consequences of Gödel's two incompleteness theorems.

Some repercussions on mathematics

The First Incompleteness Theorem

Recall the first incompleteness theorem: there exist formulae in any system capable of arithmetic which are true, but not provable. We could prove that by actually developing a formula that was true, but not provable.[1] This means that the set of provable statements in arithmetic are not identical with the set of true statements of arithmetic.

That is a result which might have already been hinted at by a variety of the Skolem-Lowen theorem applied to arithmetic. Consider the Gödel numbering process, where he showed that the set of expressions of our system could be correlated to the set of (or rather a subset of) the natural numbers. By this he has shown that the set of statements that are provable is a countable set. But consider the set of all subsets of natural numbers. This set has cardinality aleph 1, since it is the power set of the set of natural numbers.[2] Yet, in a sense one might suppose that there exists a certain number theoretical truth about each of those subsets that differentiates it from another subset of

[1] It is interesting to note that a year before Gödel published his incompleteness results, it was proved that a system of arithmetic lacking multiplication is complete, thus the threshold, "the critical mass" to make a certain formal system rich enough seems to be the multiplication operation--or in the case of second order theories Hume's Principle.

[2] Assuming for the moment the correctness of Cantor's analysis of the different types of infinity.

the natural numbers (and if it is simply the truth of an exclusion or inclusion of certain sets of numbers). Thus, if it would turn out to be possible to think of each subset of the set of natural numbers as having correlated to it a certain unique number theoretical truth, then it is easy to see that the set of provable statements in a formal system and the set of number theoretical truths is unequal, since one set is countable, whereas the other one is not. (I am aware that finitistic restrictions on what it means to be a number theoretic truth might refute such an argument as the one above.) Thus, Gödel's incompleteness theorems seem to be hinted at by other theoretical discoveries.

Be this as it may, Gödel's first incompleteness theorem was and still is a major rupture in the world of mathematics which inescapably showed the inequality of the set of true statements in an arithmetic compared with the set of provable statements:

> ...Until about fifty years ago, truth to a mathematician had been synonymous with logical proof.... [M]athematicians had operated in a fantasy world, one in which nothing was left to faith because everything could be proved to be either true or false... (Guillen, 117).

Gödel's proof showed that an axiomatic approach to number theory could not exhaust all number theoretical truths. Yet, this means that a mathematician when doing research cannot assume that a certain conjecture, which seems to hold for a finite number of cases and might in fact be true, should have a proof. But this leads us to question: just what should be understood as mathematical truth now? Gödel's proof neither tells us how many unprovable truths there are nor makes explicit what kind of faculty of truth recognition is involved in "seeing" non-provable mathematical truths. But it "introduced into the mathematical world a formal role for subjectivity, since the only possible way of avoiding unprovable truth, mathematical or otherwise, is to accept it as article of faith" (*Ibid.*). But what should be our criteria for accepting such truths?

In the case of the Gödel sentence interpreted as "I am not provable," we could see it being true by the way we constructed it. But showing that the set of provable statements is a proper subset of the set of true statements (of arithmetic) does not show us any property of the structure of those statements that lie outside of the intersection between the set of provable statements of arithmetic and the set of true statements of arithmetic. How are we to determine which statements are unprovable truths of arithmetic if we might only be able to test a statement's truths in a finite (albeit large) number of cases? Gödel himself was led to believe that only something of a platonic realism will enable us to give an adequate definition of mathematical truth:

> Classes and concepts may ... be conceived as real objects ... existing independently of our definitions and construction. It seems to me that the assumptions of such objects is quite as legitimate as the assumption of physical bodies, and there is quite as much reason to believe in their existence... (Gödel, 137).

Thus, according to Gödel, abstract objects of mathematics might be considered as having existential status comparable to, or even more perfect than that of physical objects and therefore might be perceived (by a properly trained mind) in just as lucid a way as physical objects are perceived by our senses. Yet, it barely needs to be mentioned that a "platonic realism" as regards mathematical truth is hardly undisputed.

What criteria does a mathematician have for labeling one strong, seemingly realistic intuition a lucid vision of truth, yet another insight just a misguided association? (A problem both for intuitionism as well as, I believe, for mathematical realism.) Do I really need to point out that intuition (i.e., overpowering clear visions of "the obvious") has turned out to be especially deceiving in mathematics? One needs only to consider the case of negative numbers or the existence of non-Euclidean geometries--theories violating deeply held convictions of two millennia of mathematical tradition--in order to see how much intuition of the "obvious" can lead researchers astray.[3] Furthermore, what criteria do we have in order to accept the lucid vision of one mathematician contradicting another mathematician's most clear vision of a "mathematical truth"?

To these questions Gödel replied that a thoroughgoing philosophical analysis of mathematical ontology will yield sufficient answers, presuming a realistic position. Be that as it may, the foundation assumption that there exist mathematical, disembodied, eternal forms for us to be discovered as Columbus discovered America (Nagel, 99) has been an upsetting prerequisite for many mathematicians. They have attempted to come to terms with Gödel's results about the incongruity of mathematical truth with mathematical proof in different ways. The formalists, as they are usually alluded to, "..believe that mathematics is purely an invention of the human mind ... Proving an hypothesis only means that it is a successful invention, much like an airplane that actually flies..." (Guillen, 123). Mathematics thus becomes comparable to the empirical sciences, i.e., "a hypothesis is tentatively declared to be true if it is the simplest available explanation of the evidence" (Guillen, 121).

It must be realized that this stance opens another Pandora's box: we have to design guidelines of just what is meant by empirical adequacy for a probable conjecture to become a tentative truth. How many instances of being successful on a given test set makes a conjecture a tentative truth? How many years of resisting proof or disproof by the most eminent scholars in the field justify considering a conjecture the instance of an unprovable truth? It seems to me that no rigorous answer can be given to these questions--the infamous paradox of the heap seems to loom over any definite quantifying reply. For example, why should 200 years of evidence in favor of Goldbach's conjecture, rather than 150 or 250 years, render the hypothesis a tentative truth?! A definite answer seems to draw a more or less arbitrary line. This is exactly what seems to unify the notion of unprovable truth in mathematics with the notion of scientific truth (i.e., how much and what kind of empirical data should render a certain hypothesis a tentative truth). This similarity should allow for some of the techniques and concepts of scientific truth to

[3] Descartes still thought that the concept of a negative number was utterly meaningless.

translate into mathematical practice. Yet, it must be stressed here that Gödel's proof did not disown mathematics of its crown--the irrefutable proof.

It is sometimes pointed out that researchers in mathematics seem to ignore Gödel's proof and spend their time attempting to proof conjectures instead of trying to amass a large body of true sample cases. But that seems an exaggeration of the consequences of Gödel's incompleteness results! There is simply no tag to a certain conjecture that tells a researcher that this particular object is an unprovable truth or just an exceptionally hard object to prove (or disprove). As I pointed out in the beginning, Gödel's proof does not specify at all what characteristics unprovable truths should have, nor does the proof state how many unprovable sentences there are, beyond the obvious statement that there is at least one such truth. Thus, the assumption that a particular object is provable appears to be the safest assumption when treating a mathematical object, since the only way to decide assuredly whether a certain conjecture is an unprovable truth is to either prove or disprove it.[4] However, even if proof fails, we have added "empirical evidence" for the object's special status, which might at some time become so overwhelming that it is difficult to doubt either it's truth or its unprovability. Thus, assuming that a conjecture of uncertain status is provable, and proceeding accordingly to attempt proof or disproof, appears to me the most fruitful assumption which leads in *any* outcome to an increase of information about the conjecture in question.

In summary, it must be said that the first incompleteness theorem represents a major stir in the field of mathematics. The waves thus created have not abated. On the contrary, ever since Gödel's proof, the definition of mathematical truth has been heatedly disputed, allowing more easily for the introduction of the subjective, "extra-logical" (or rather "extra-formal") truth recognition abilities. It seems ironic that Hilbert's program, which sought to establish a secure, formal foundation for mathematics, should have in the end led to the demise of one of mathematics' most deeply entrenched postulates: namely, that truth equals proof.

Let us now consider some repercussions of Gödel's second incompleteness theorem on mathematics.

<u>The Second Incompleteness Theorem</u>

The proof of the first incompleteness result depended on the assumption that the system of arithmetic in which we conduct our proof is consistent. We are then led to ask whether we can prove the consistency of the system under question within our system. Yet, recall Gödel's second incompleteness result: He showed that it is impossible to prove the consistency of a formal system powerful enough to express arithmetic purely within the system itself.

[4] "..repeated failure to construct a proof does not mean that none can be found any more that repeated failure to find a cure for the common cold establishes beyond doubt that mankind will forever suffer from running noses..." (Nagel, 10).

The question of consistency became especially pressing when the perception of mathematics as the "science of quantity" gave way to a new way of looking at mathematics:

> ... It became evident that mathematics is simply the discipline *par excellence* that draws the conclusions logically implied by any given set of axioms or postulates. In fact, it came to be acknowledged that the validity of a mathematical inference in no sense depends upon any special meaning that may be associated with the terms or expressions contained in the postulates... (Nagel, 11).

The shift of the view of mathematics was in part caused by the discovery of non-Euclidean geometries. A method for proving that a given set of axioms would not lead to contradictory theorems had to be established. In trying to prove Euclidean geometry consistent, Hilbert reduced Euclidean geometry to an algebraic model. But the reductionistic route to proving consistency simply passes on the buck, since it proves the consistency of the reduced system only if the system that it is reduced to is consistent.[5] This difficulty, among other things, led Hilbert to include a call to establish a proof of the "absolute consistency" of a system of arithmetic in his above-mentioned reformation program for mathematics.

An absolute proof of consistency is a proof "...by which the consistency of the system [can] be established without assuming the consistency of some other system..." (Nagel, 26). In other words, it is a proof which does not try to establish its result by standing on the shoulders of another system, however obvious that system might be. The difference between relative (reductionistic) and absolute proof of the consistency of a system is in some way similar to the difference between hypothetical imperative and categorical imperative: the former, however obvious, can only be true if its antecedent condition, however obvious, can be shown to hold. The latter turns out true no matter what; it is independent of other postulates and, so to speak, "grows out of itself" (at least, if one is a Kantian).

Can there be an absolute proof of the consistency of arithmetic? As Gödel's proof shows it is highly unlikely that such a proof should be available. It excludes an absolute proof of the consistency of arithmetic within the very formal systems of arithmetic. However, this result does not exclude absolute proofs of arithmetic *per se*. As Nagel points out:

> ... The possibility of constructing a finitistic absolute proof of consistency for arithmetic is not excluded by Gödel's results. Gödel showed that no such proof is possible that can be represented within arithmetic. His argument does not eliminate the possibility of strictly finitistic proofs that cannot be represented within arithmetic. But no

[5] Such proofs are alluded to as relative proofs of consistency.

> one today appears to have a clear idea of what a finitistic proof would be like that is not capable of formulation within arithmetic... (Nagel, 98).

But what should a consistency proof about a system look like, that does not include the very operations (or isomorphic manipulation rules thereof) that characterize the system of arithmetic? The unlikeness of such proof to exist stems from the reflection that in order to show arithmetic consistent, we will somehow have to allude to multiplication and addition. But it seems highly unlikely that a (first-order) formal system can be constructed that alludes to these operations without being itself thereby powerful enough to express these operations in some way or other.

However, notice again that we seem to be left in a state of suspension where we can only make a probable conjecture. Gödel once again has left our judgment about an important part of mathematics, the absolute consistency of certain formal systems, in a limbo whose consequences as yet still have to be fathomed.[6] But the consistency proofs of formal mathematical systems are extremely important both for logicism and formalism. In both endeavors, an attempt is made to reduce parts (or all) of mathematics to formal systems. But when such reduction is done, it is important to establish that the resulting system is consistent, so as not to allow us to derive anything we like in such a system. But a proof of this kind within the system of reduction is precisely what is excluded by Gödel's second incompleteness theorem. All the logicist or formalist can do is either to engage in a relative consistency proof (where he/she proves the consistency of his/her system assuming the consistency of some stronger formal system), or to attempt to find a way to construct consistency proofs for a formal mathematical system without including the operations of arithmetic in his/her system of the consistency proof. Both attempts to show consistency seem to me problematic, for they either assume what is to proven in a stronger system, or seem to attempt to prove the effect of arithmetical operations without wanting to use them in their consistency proof. Gödel's proof thus poses important questions both for the foundations, as well as the practice, of mathematics. However, the consequences of Gödel's proof seem to go beyond the realm of mathematics, as we will see below.

A Refutation of artificial intelligence--Or Is It?

Gödel's results have been claimed to be relevant for the field of artificial intelligence (AI). In the following section, I will attempt to show that some of the criticisms of artificial intelligence based on Gödel's results are not well founded. To begin, let us consider just how the results of Gödel come to be an apparent difficulty for the prospect of artificial intelligence.

[6] Using a higher order system and non-finitistic methods, the consistency of arithmetic can be proved, but this is of course not an absolute, but relative, proof of arithmetic's consistency.

We should first consider just what is meant by the fashionable term, "artificial intelligence." Unfortunately, the workers of the field have not completely agreed on a crisp specification of just what is meant by saying that an object is artificially intelligent. At the heart of artificial intelligence seems to lie the hope either to create an entity that is capable of rational actions in a wide variety of domains (i.e., we would like the machine to be able to do what we can do, or do it better), or to create an entity that is in some ways an equivalent of the human mind (whatever that is), i.e., a machine that has a mind like ours. It should be obvious that the fulfillment of the second goal should lead to the fulfillment of the first (i.e., how else but by the behavior can we test whether the structures that we imputed into the machine as capturing mindfulness really produce the effects of having a mind in the machine!). However, it is not necessarily the case that the fulfillment of the first goal (i.e., an entity that "acts like us"[7]) must lead to the fulfillment of the second goal. This is the case since a structure unlike that of the human mind might cause an agent to act like a rational, general purpose intelligent agent. But consider that in either case, the means by which to arrive at either one of the goals seems to be the creation of a computational model (or implementation). Yet, what is a computational model (of the mind, or of a rationally acting agent)?

According to Church's thesis, everything that is an algorithm is Turing Machine computable. But it is by the specification of appropriate algorithms that researchers in AI hope to fulfill either one of the goals mentioned above. By Church's thesis, this means there exists a Turing Machine equivalent to the collection of algorithms that are sought. Such a line of research implicitly assumes the Church-Turing thesis which postulates that "...Mental processes of any sort can be simulated by a [Turing Machine]..." (Hofstadter, 578). In fact, some adherents of AI might go farther and assume that it is possible for a certain Turing Machine to emulate, not just simulate, any mental process, i.e., to be a (human like) mind, not to simulate one.[8] But a Turing Machine is the specification of a formal system obviously powerful enough to express arithmetic.[9]

Yet this implies that Gödel's results seem relevant, i.e., given any Turing Machine capable of expressing arithmetic, there exists a formula that is true but not provable within the system specified by the Turing Machine. Can you see how the argument against the possibility of a Turing Machine capturing mindfulness takes off? Hofstadter gives a good summary:

> ..Rigid internal codes entirely rule computers and robots; ergo..
>
> Computers are isomorphic to formal systems..

[7] I do not feel it possible to conceive of an agent utterly different from our recognized way off displaying intelligence.

[8] It is unclear where the distinction between simulation and emulation breaks down.

[9] Turing machine computability has been proved to be equivalent to Abacus computability.

> Any Computer which wants to be as smart as we are has got to be able to do number theory as well as we, so..
>
> Among other things, it has to be able to do primitive recursive arithmetic. But for this very reason
>
> It is vulnerable to the Gödelian "hook," which implies that
>
> We, with our human intelligence can concoct a certain statement of number theory which is true, but the computer is blind to that statement's truth (i.e., will never print it out), precisely because of Gödel's boomeranging argument. This implies that there is one thing which computers just cannot be programmed to do, but which we can do. So we are smarter... (Hofstadter, 472).

At first this argument seems extremely conclusive. But let us notice at the start that this line of reasoning is a problem to AI only if we make two assumptions: (1) that a machine used in AI ought to be a device that is capable of expressing arithmetic, and (2) that it is a fundamental goal (even a necessity) for the field of AI to create a machine equivalent to human beings in power of mathematical truth recognition. Point (1) can scarcely be disputed, since the design of agents in AI is a creation of a computational artifact. It is difficult to see how any mechanistic system, supposedly powerful enough to display rational behavior of a complex sort, would not be captured by a formalistic description powerful enough to express arithmetic. In Hofstadter's words, there seems no alternative to creating a system with a certain critical mass of expressive power when one wants to create a system powerful enough to act reasonably.

However, point (2) is not unanimously acknowledged as a requisite for AI. In fact, most current research aims to produce not a Golem, but rather an artifact acting intelligently in a restricted problem domain. If it is granted that this is what all of AI should be about, Gödel's proof would hardly be a challenge. We would accept the limitation and move on.

Yet, as I insinuated above, there seems to be a deeply hidden dream at the heart of AI: either to produce a mind like ours by mechanistic means, or to create an artifact that acts like a human being in a general problem domain.[10] It is this underlying theme whose realization seems denied by the kind of reasoning of the meaning of Gödel's proof above. Let us look at such a conclusion:

[10] Without making the ontological claim that the artifact must possess a mind structured like ours. Of course, it seems difficult to imagine that it is possible to create such an artifact, the "mind" of which is wholly dissimilar from the human mind but which exhibits traits like communication skills, maybe emotion, and certain problem-solving techniques similar to human beings.

> .... We are trying to produce a model of the mind which is mechanical--which is essentially "dead"--but our mind, being in fact "Alive," can always go one better than any formal, ossified, dead system can. Thanks to Gödel's theorem, the mind always [my emphasis] has the last word... (Lucas, quoted in Hofstadter, 472).

But is it really an insurmountable challenge to the Church-Turing hypothesis? Will it really deny conclusively the faintest possibility of creating a mechanistic mind? We have various ways of responding.

Let us reconsider the outcome of Gödel's (first) incompleteness theorem. In plain English, it merely states that there will always be an unprovable truth in any formal system powerful enough to express arithmetic. Gödel's results do not state that it is always possible for a human mind to discern such a truth! For example, we know that there are an infinite number of primes. Yet any number so long that no human being will ever be able to complete reading it in her/his lifetime is surely a number for which no human can decide whether or not it is a prime number. But such numbers exist (in fact, an infinite amount of them exists). Thus, the knowledge that a certain fact must hold in general (in our case for the domain of natural numbers) does not entail that a human being is always able to prove its truth for arbitrary instances of the problem domain.

But this is exactly what seems to be assumed by Lucas and his skeptical soul mates, i.e., a variation on Lucas' last sentence seems to make their arguments definitive: "...the mind can always go one better than any formal [...] system...". But that seems plainly wrong. It is legitimate to keep amending the formal system with proper axioms capturing Gödel truth sentences (i.e., we would as a first step include "This sentence is not provable" in our initial formal system as a truth), so that the system recognizes the sentence(s) that were formerly shown to be true, but unprovable. It follows from Gödel's results that such a system in turn would include a formula that is true, but not provable. But there certainly exist systems for which the construction of Gödel's proof would take a human, say, 10,000 years, because all the necessary formulae are (although finite) so large that it might take 100 years to write one of them down. In fact, it is perfectly clear that there exist systems which are proper first-order theories, but for which there might be proper axiom specifications that would take an arbitrary, denumerable amount of time to write down, whatever (finite) speed is used to write it (or read it, for that matter). Thus, there will always exist proper first order theories for which it is impossible that a human could ever complete the application of certain, extremely large axiom schemata. But in the same vein there exist formal systems that are so complex that no human being will be able to apply their axioms successfully, much less being able to apply Gödel's method of proof. Thus, it is simply untrue that the human mind can always show, or "Outgödel," a given formal system. As Hofstadter points out:

> ...In fact as the formal systems (or programs) escalate in complexity, our own ability to "Gödelize" will eventually waver.... We do not have any algorithmic way of describing how to perform it. If we can't tell explicitly what is involved in applying the Gödel method in

> all cases, then for each of us there will eventually come some case so complicated that we simply can't figure out how to apply it.... Of course, this borderline of one's abilities will be somewhat ill-defined, just as is the borderline of weights which one can pick up off the ground. While on some days you may not be able to pick up a 250-pound object, on other days maybe you can. Nevertheless, there are no days whatsoever on which you can pick up a 250-ton object. And in this sense, though everyone's Gödelization threshold is vague, for each person, there are systems which lie far beyond his ability to Gödelize... (Hofstadter, 475).

It might be objected that this misses the point. We are not justified in shifting the discussion to the current limitations on the ability of humans to grasp formal complexity, since arbitrary biological constraints might be overcome, thus making the above rebuttal only a somewhat (if highly) probable claim. Gödel's proof points instead, it is argued, to the fact that there is an essential difference between truth sensed by a human mind as compared to truth that a formal system might be able to prove within itself. First of all, it is not clear how this criticism seems to apply. The difference alluded to is one in recognizing a certain truth about certain formal systems, while the formal system itself is not able to recognize that fact [the non-provable true formula(e)]. But no one is claiming that a spartanic system of arithmetic is a human mind. The above injunction, that a human mind can find a truth about certain formal systems while that formal system cannot, does not imply that the human mind could not itself be adequately described by a complicated formal system. Thus, pointing to the fact that the human mind sees truth in some formal systems, whereas the formal system does not, seems only to imply that the mind is certainly a stronger system than the one it can see truth in. It does not imply that the source for seeing such a truth is something that in principle lies beyond the description of a sufficiently complex formal system. To assume this is simply begging the question, since the burden of proof in this specific instance rests with the proponents who claim that Gödel's proof refutes the endeavor of artificial intelligence:

An interesting fact is that also humans are liable to be "Gödelized":

> ...C. H. Whitely ... proposed the sentence, "Lucas cannot consistently assert this sentence." If you think about it you will see that (1) it is true and yet (2) Lucas cannot consistently assert it. So Lucas is also "incomplete" with respect to the truths about the world. The way in which he mirrors the world in his brain structures prevents him from simultaneously being "consistent" and asserting that true sentence. But Lucas is no more vulnerable to us than any of us. He is just on par with a sophisticated formal system... (Hofstadter, 477).

Thus, in a sense, even human beings are necessarily incomplete systems when viewed from the standpoint of logic. It therefore appears to be an unfair requirement to demand that certain AI artifacts should be complete with respect to the truths of the world, whereas human beings are not.

But this does not exhaust our rejoinders to arguments *à la* Lucas. One might in fact reply that a dismissal of AI on the basis of Gödel's proof might be completely missing the point. Neither AI nor human conduct of thought seems to be based on a notion of proofs. Humans approximate; they use heuristics (probable guidelines) even in proving a certain thing in a formal system.[11] Human beings are dramatically fallible in doing proof, and it seems a fair assumption that not all--maybe not even the majority--of today's human (adult) minds in the world will be capable of performing Gödel's proof; yet one does not doubt the concert violinist's intelligence. It might be said that this does not represent a particularly sound argument, since a human mind might in principle be capable of understanding and producing Gödel's proof: environmental influences, which are now so accidental that only a few elect people understand the technicalities involved, might in principle be so generated that most or all fully accountable human beings, given some training time, will be able to "Gödelize" (as if they had nothing better to do). Such an assumption seems stark, but we must not lose ourselves in slippery-slope arguments about the portion of human minds that might be able to comprehend and use Gödel's proof technique. Instead, we assume the contrary: namely, that there exists at least one human being who is treated as a person with full-fledged rights, but who--even given proper training units--lacks the ability to comprehend and use the Gödel technique.[12] Given that such a person is very likely to exist, it seems certainly true that we do not come to withhold personhood status from this human simply because of the failure of successfully utilizing a few esoteric math procedures. Now, if it is granted that at least one such individual exists, it seems an unjustified double standard of expectation to make the prospects of AI dependent on producing minds that recognize the Gödel proof procedures, whereas we do not have such a prerequisite for granting personhood status to some human being. If we still withhold personhood status, it must be for the lack of other characteristics which seem essential, but are rarely spelled out by Lucas and followers who assume that a failure to meet the Gödelean truth recognition is reason enough to discredit AI. In fact, the critique goes farther: as pointed out above, human beings seem to use heuristic guidelines to steer their behavior in the world.

> The question turns on whether the fact that a human mathematician can always recognize as true propositions that cannot be proven to be true within a given formal system shows that human beings cannot be modeled by an information-processing model which is necessarily a

[11] It suffices to review the many books which promulgate this or that collection of heuristics to become more successful in doing mathematical proofs.

[12] This seems at least a very probable assumption heeding the many differences in innate potentialities and capabilities.

> formal system. But such an argument misses the point. Even if AI did produce an information processing model of a mathematician, that model would be able to see that a specific formula was true by means of calculations based on its heuristic rules. Of course, the heuristic calculation could itself be viewed as a proof that certain conclusions follow from certain premises, but these premises would be formulae describing what the mathematician perceived, believed, remembered, etc., and the 'Conclusions' would be what he would say or surmise, etc.--obviously not the premises and conclusions of an acceptable formal proof of the original formula... (Dreyfus, 345).

Dreyfus thus makes an interesting claim here: AI might be able to adduce approximation rules, some form of heuristics modeling the "extra-logical" truth recognition abilities that seem at first sight involved in finding "non-provable" truths. Here we have another way to avoid the mouse trap for AI set up by Gödel's proof. We simply try to include a cognitive model of non-formal truth recognition capabilities in our computational model, therefore producing an agent endowed with a certain "self-introspection." Obviously the addition of such a cognitive model is no easy step, but it seems hardly precluded by Gödel's first and second incompleteness theorem: both theorems apply to the notion of truth and provability, whereas heuristic rules are not (at least in the apparent sense) a proof technique.

To me, all the above-mentioned replies to a critique of AI based on an argument comparable to Lucas' seem to insinuate that the initial reaction that Gödel's proof refutes the feasibility of artificial intelligence is quite disputable. Gödel's proof does not seem to doom the secret hopes of artificial intelligence, at least in the way I have presented it here.

Conclusion

We have come a long way. There are a few points I would like you to carry away from this essay. First of all, I have tried to give a short overview of Gödel's proof which constructs a sentence of the form "I am unprovable" in arithmetic. The arithmetic equivalent of the above sentence resulted in the two incompleteness theorems of Gödel: that there are truths within arithmetic that are not provable, and that we cannot prove the consistency of a particular system of arithmetic within that very system itself. These results have uprooted deep convictions in the field of mathematics and have further entrenched the gulf between metaphysical approaches to the foundations of mathematics. Mathematics has in part taken the form of an empirical study because of Gödel's results about mathematical truth. The importance of Gödel's results has not yet been fathomed and will presumably be a point of discontent and search for disciplinary standards in the near future. However, the proof has not upheld the hopes of the critics of AI, namely that it would dismantle the deep assumptions that a mind could be adequately described by a formal system. It appears to me that the consequences of Gödel's proof for mathematics are as substantial as they are minimal for artificial intelligence.

Bibliography

Crossley, J. N., *What Is Mathematical Logic?* New York: Dover Publications, Inc., 1990.

Dreyfus, Hubert L., *What Computers Can't Do.* New York: Harper Colophon Books, 1979.

Gödel, Kurt, "Russell's Mathematical Logic," pp. 123-154 in *The Philosophy of Bertrand Russell,* edited by Paul A. Schilpp. Evanston: Northwestern Univ. Press, 1944.

Guillen, Michael, *Bridges to Infinity.* Los Angeles: J. P. Tarcher, Inc., 1983.

Hofstadter, Douglas R., *Gödel, Escher, Bach.* New York: Vintage Books, 1980.

Linz, Peter, *An Introduction to Formal Languages and Automata.* Lexington, MA: D. C. Heath Company, 1996.

Nagel, Ernest, and James R. Newman, *Gödel's Proof.* New York: New York Univ. Press, 1958.