Appeared in: *Revue Internationale de Philosophie* 59 No 4 (a special issue on Gödel)

On the Philosophical Relevance of Gödel's Incompleteness Theorems

Panu Raatikainen

Gödel began his 1951 Gibbs Lecture by stating: "Research in the foundations of mathematics during the past few decades has produced some results which seem to me of interest, not only in themselves, but also with regard to their implications for the traditional philosophical problems about the nature of mathematics." (Gödel 1951) Gödel is referring here especially to his own incompleteness theorems (Gödel 1931). Gödel's first incompleteness theorem (as improved by Rosser (1936)) says that for any consistent formalized system *F,* which contains elementary arithmetic, there exists a sentence G_F of the language of the system which is true but unprovable in that system. Gödel's second incompleteness theorem states that no consistent formal system can prove its own consistency.[1]

These results are unquestionably among the most philosophically important logico-mathematical discoveries ever made. However, there is also ample misunderstanding and confusion surrounding them. The aim of this paper is to review and evaluate various philosophical interpretations of Gödel's theorems and their consequences, as well as to clarify some confusions.

The fate of Hilbert's program

It is widely thought that Gödel's theorems gave a death blow to Hilbert's program. Whether Gödel's theorems really demonstrated that it is impossible to carry out Hilbert's program is controversial. This is partly because there is not complete clarity as to what exactly constitutes Hilbert's program, and what views are truly essential for it. Furthermore, some of Hilbert's key concepts are somewhat vague. Nevertheless, I think that there are good reasons to think that Hilbert's mature program of the 1920s was, in its original form and in its full generality, refuted by Gödel's theorems.[2]

[1] It should be noted that in their full generality, Gödel's theorems presuppose a mathematical explication of the intuitive notion of effective calculability or decidability, which was provided by Turing.

[2] I have argued for my own interpretation of Hilbert's program in detail in Raatikainen (2003a).

Hilbert made two fundamental distinctions. First, he distinguished between unproblematic and contentful finitistic mathematics and contentless infinistic mathematics. It is now usual to assume that finitistic mathematics is essentially captured by Primitive Recursive Arithmetic PRA. Second, Hilbert made the distinction between real sentences and ideal sentences. He thought that only real sentences are meaningful and have real content. These are roughly quantifier-free formulas preceded by one or more universal quantifiers. All the other sentences are ideal sentences, meaningless strings of symbols which complete and simplify formalism and which make the use of classical logic possible.

Hilbert's program was planned to proceed as follows: first, all of infinistic mathematics was to be formalized; next, one should, using only restricted and uncontroversial finitistic mathematics, prove the consistency of this comprehensive system; moreover, one should show that infinistic mathematics would never prove meaningful real sentences that were unprovable by finitistic mathematics. This would guarantee the safety and reliability of using infinitary methods in mathematics which, after set-theoretical paradoxes, had been questioned by many.

Under the natural assumption that finitistic mathematics is recursively axiomatizable, Gödel's results establish that it is impossible to carry out Hilbert's program in its original form — even if one does not need to formalize at once the whole mathematical truth (which is trivially impossible by Gödel's theorems) but just some existing piece of infinistic mathematics (say, second-order arithmetic). By Gödel's theorems, a strong infinistic theory always proves 'real sentences' which are unprovable by finitistic mathematics. Understood in this way, Hilbert's program was truly refuted by Gödel's theorems (see Raatikainen 2003a).

Conventionalism, syntax and consistency

Although Gödel was originally a member of the Vienna Circle, and his views on the philosophy of mathematics, as they developed, were clearly at odds with those of the logical positivists, Gödel did not much comment on this conflict in his publications. Gödel was, nonetheless, preparing a contribution to the Carnap-volume of Schilpp's *Library of Living Philosophers*, but he was unsatisfied with his manuscript, and finally decided not to publish it. In this manuscript (Gödel 1953/9), Gödel develops a conclusive argument against conventionalism, that is, he attacks the view "which interprets mathematical propositions as expressing solely certain aspects of syntactical (or linguistic) conventions". He mentions Carnap, Schlick and Hahn as advocates of this position.

According to Gödel, a rule about the truth of sentences can be called syntactical only if it is clear from its formulation, or if it somehow can be known beforehand, that it does not

imply the truth or falsehood of any 'factual sentence' or 'proposition expressing an empirical fact'. But, so the argument continued, this requirement would be met only if the rule of syntax is consistent, since otherwise the rule would imply all sentences, including the factual ones. Therefore, by Gödel's second theorem, the mathematics not captured by the rule in question must be invoked in order to legitimize the rule, and thereby the claim that mathematics is solely a result of syntactical rules is contradicted.

Now although Gödel addressed this paper especially to Carnap, he did not pay close attention to the possible differences between the members of the Vienna Circle. Consequently, it has been argued by Goldfarb and Ricketts that Carnap's radical and sophisticated variant of conventionalism is in fact immune to this sort of direct refutation (see Goldfarb and Ricketts 1992, Ricketts 1995, Goldfarb 1995). This interpretation has been, in turn, questioned by Crocco (2003). However, be that as it may, Gödel's argument is in any case fatal to the more standard forms of conventionalism, such as those of Schlick and Hahn. And this is certainly already very interesting in itself.

Self-evident and analytical truths

One can also provide more general epistemological interpretations of Gödel's theorems. Quine and Ullian (1978), for example, consider both traditional rationalist philosophers who believed that whatever was true could in principle be proved from self-evident beginnings by self-evident steps, and the "less sanguine" ones who argued that whatever was true could be proved by self-evident steps from two-fold beginnings: self-evident truths and observation. Contrary to both schools, Quine and Ullian point out that even the truths of elementary number theory are presumably not in general derivable by self-evident steps from self-evident truths: "We owe this insight to Gödel's theorem, which was not known to the old-time philosophers." (Quine & Ullian 1978, p. 64–65.)

Hilary Putnam (1975) submits that the statements that can be proved from axioms which are evident to us can only be recursively enumerable — unless an infinite number of irreducibly different principles are at least potentially evident to the human mind, a supposition he finds "quite incredible". Hence, by Gödel's theorems, some truths of elementary number theory are not provable from evident axioms. Putnam continues that even if it were the case that all the axioms we use in mathematics are 'analytic', as some philosophers have claimed (which, he adds, has never been shown), it would not follow that all truths of mathematics are analytic. Putnam concludes that if the analytic sentences are all consequences of some finite list of Meaning Postulates, then it is a consequence of Gödel's theorem that there must be synthetic truths in mathematics (Putnam 1975).

In fact, Gödel himself made remarks in a very similar spirit. That is, Gödel first noted that 'analyticity' may be understood in different ways. One alternative is the purely formal sense that the terms occurring can be defined (either explicitly, or by rules for eliminating them from sentences containing them) in such a way that the axioms and theorems become special cases of the law of identity and disprovable propositions become negations of this law. Gödel concluded that in this sense of 'analyticity', even the theory of integers is demonstrably non-analytic, provided that one requires of the rules of elimination that they allow one to actually carry out the elimination in a finite number of steps in each case. For this would imply the existence of a decision procedure for all arithmetical propositions (Gödel 1944).

Intuitionism, truth and provability

The relation of Gödel's theorems to intuitionism is less straightforward. On the one hand, they seemed to confirm the intuitionists' misgivings about formalism. On the other hand, they underline the rather abstract nature of the intuitionistic notion of provability, with which intuitionists equate truth. For as a consequence of Gödel's theorems, truth cannot be equated with provability in any effectively axiomatizable theory.

In Gödel's own mind, at least, this is quite a serious drawback (see Gödel 1933, 1941); he complained that the intuitionistic notions of provability and constructivity are vague and indefinite and lack complete perspicuity and clarity. It cannot be understood in the sense of 'derivation in a definite formal system', since for this notion, the axioms of intuitionistic logic would not hold. So the notion of derivation or of proof must be therefore taken in its intuitive meaning as something directly given by intuition, without any further explanation. According to Gödel, this notion of an intuitionistically correct proof or constructive proof lacks the desirable precision.

It is indeed arguable that Gödel's theorems pose, for many variants of intuitionism, a much more serious challenge than has been realized or admitted by intuitionist philosophers of mathematics (see Raatikainen 2004). That is, intuitionists often emphasize that one should recognize a proof when one sees one. Put differently, proofs are understood as beginning with immediate truths, and continuing with immediate inference. But given Gödel's theorems, it is then hard to hold the intuitionistic equation of truth with probability. The problem is the same as found above with self-evident truths: is it really plausible to assume that there are infinitely many[3] irreducibly different principles which are self-evident to the human mind?

[3] actually even much worse: Gödel's technique entails that the set of self-evident truths would not be even arithmetically definable, in other words, not only non-recursive but nowhere in the arithmetical hierarchy (assuming it is legitimate to use these notions), but it would be at least as complex and abstract as the set of

Let us also note in this context an application of Gödel's theorem by Putnam (1967). Thus, consider the following two principles which, in Putnam's words, "many people seem to accept":[4]
(i) Even if some arithmetical (or set-theoretical) statements have no truth value, still to say that any arithmetical (or set-theoretical) statement that it has (or lacks) a truth value is itself always either true or false (i.e. the statements either has a truth value or it does not).
(ii) All and only decidable statements have a truth value.
Putnam shows that these two principles are together inconsistent, by applying Gödel's first theorem.

Logicism and Gödel's theorems

There has been some dispute on the issue as to whether Gödel's theorems conclusively refute logicism, that is, the claim that mathematics can be reduced to logic, as endorsed, for instance, by Frege and Russell. Obviously this issue depends heavily on how one understands the essence of logicism. Clearly Gödel's theorems show that all arithmetical truths are not reducible to the standard first-order logic, or indeed, to any recursively axiomatizable system. On the other hand, one may restrict the logicist thesis to some class of mathematical truths (such as known truths, or humanly knowable ones), and/or extend the scope of logic. There is, though, the threat that the issue becomes trivial or wholly verbal.

Henkin (1962) and Musgrave (1977), for example, state that Gödel's results effectively destroy classical logicism (see also the comments by Quine, Ullian and Putnam concerning self evidence of the mathematical truths found below). Sternfeld (1976) and Rodríguez-Consuegra (1993), on the other hand, argue that it is possible to defend logicism even after Gödel's theorems. Sternfeld and Rodríguez-Consuegra appeal to the fact that Gödel's theorems do not provide an absolutely undecidable statement, but only a relative one. This is certainly true. Yet this defense apparently collapses logicism into the view that every mathematical truth is derivable in some formal system. This, however, makes the thesis completely trivial. Furthermore, would this not imply that not only mathematics but also all empirical facts are 'logically true'?

Geoffrey Hellman (Hellman 1981, see also Reinhard 1985) has analyzed the bearing of Gödel's theorems on logicism in more detail. Hellman focuses only on the thesis that *knowable* mathematical truth can be identified with derivability in some formal

classical truths (of arithmetic).

[4] though, it should be added, this is not the view of standard intuitionism.

system. Logicism so understood cannot be directly refuted by Gödel's first theorem. Hellman subsequently gives a considerably more complicated argument which leans on Gödel's second theorem, and breaks the argument down into two cases. First, he concludes that no finitely axiomatizable logicist system exists. Second, he considers non-finitely axiomatizable systems, and here the claim is weaker: such logicist systems may exist, but Gödel's second theorem prohibits our being able to know of any particular system that it is one of them.[5] Hellman's argument has the advantage of not depending on any particular restrictive way of drawing the controversial line between logic and non-logic.

Incompleteness and Algorithmic Complexity

In the past few of decades, certain variants of incompleteness results, together with their ambitious philosophical interpretations, by the American computer scientist Gregory Chaitin, have received considerable attention. It has often been suggested that Chaitin's results are fundamental and dramatic extension of Gödel's results, or even the strongest possible version of an incompleteness theorem, and that they shed new light on the incompleteness phenomenon and explain why it really occurs.[6]

Chaitin's results emerge from the theory of algorithmic complexity or program-size complexity (also known as "Kolmogorov complexity"). In fact, Chaitin himself was one of the founders of that theory. The algorithmic complexity, or the program-size complexity, of a number or a string, refers to the length of the shortest program which generates the number or string and halts. A finite string is referred to as random, or irregular, if its complexity is approximately equal to its length. Further, an infinite sequence is called random if, roughly, all its finite initial segments are random (this requires qualifications). It has been also proposed, for somewhat unclear and confused reasons, that algorithmic complexity provides a good measure of the information content of a string of symbols. Consequently, the whole field is often called Algorithmic Information Theory.

It was known from the beginning that program-size complexity is undecidable. However, in the early 1970s, Chaitin observed that it has a peculiar property: Although there are strings with arbitrarily large program-size complexity, for any consistent mathematical axiom system, there is a finite limit *c* such that in that system, one cannot prove that any particular string has a program-size complexity larger than *c* (Chaitin 1974, 1975). Later, Chaitin attempted to extend the complexity-theoretic approach in order to obtain "the strongest possible version of Gödel's incompleteness theorem" (Chaitin 1987b, p. v). For

[5] The latter, weaker conclusion resembles the conclusions drawn by Benacerraf as well as Gödel's related conclusions; see below.

[6] For an in-depth criticism of these interpretations, see Raatikainen 1998, 2000; cf. also Raatikainen 2002.

this purpose, Chaitin has defined a specific infinite random sequence Ω ("the halting probability"), and then showed that no formal system F can determine but finitely many digits of Ω (Chaitin 1987a, 1987b).

Chaitin's results are not without interest, but one should not exaggerate their power or relevance. Also, the popular explanations and philosophical interpretations arising from these results are largely unsupported by facts. Both Chaitin's incompleteness results exhibit a finite limit of provability; Chaitin maintains that these limits, for a given axiom system, are moreover determined by the algorithmic complexity of the axiom system. Yet this is based on confusions and is just not true. In fact, there is no correspondence between the two. One can have extremely complex but very weak systems with a small limit and quite simple but very strong systems with a much larger limit.

Chaitin has further interpreted his results as showing that the incompleteness phenomenon occurs because undecidable sentences "contain too much information", that is, more than the axioms of the theory, and stated that this is the ultimate explanation of incompleteness. Yet, this is not true in general. It is false for both 'algorithmic information' (i.e. program-size complexity) and for the intuitive common-sense notion of informativeness. There is no correspondence between the complexity of axioms and the complexity of the undecidable sentences. It is wholly possible to have an extremely complex axiom system with a strikingly simple sentence which is undecidable in the system. And intuitively speaking, the Gödel sentences, which are just particular universal statements, are usually much less informative (in the intuitive sense) than the formalized theories from which they are independent. While it is true that they contain information the system does not contain, it does not follow that they contain more information than the system.

Chaitin's results are also not "the strongest possible" incompleteness and undecidability results. In a sense, Gödel's and Turing's classical results are stronger than Chaitin's earlier (1974) incompleteness result, for the former provide a *m*-complete set, whereas Chaitin's result does not, and the undecidable set of the latter is *m*-reducible to the undecidable sets of the former but not *vice versa*. Nor is Chaitin's Ω the extreme of undecidability, as it has been sometimes called. In fact, there are certain in a definite sense more strongly undecidable arithmetical problems which are in addition much more natural (see Raatikainen 2000; a particularly nice example, in terms of ordinary number theory, can be found in Raatikainen 2003b).

Why is the Gödel sentence true?

Apparently people have no difficulties in understanding the idea that a formal system leaves some sentences undecided. Nevertheless, confusion surrounds the reasons for

holding, in Gödel's first theorem, the undecided Gödel sentence to be true. Some apparently think that humans can intuitively see that it is true, perhaps because "it says of itself that it is unprovable". Others assume that we somehow check that it holds in the standard model of arithmetic. Still others think that the question of its truth is meaningful only when understood in terms of provability in some other, stronger system. All such views are problematic and irrelevant. Let us attempt to understand more clearly what the real state of affairs is.

The structure of Gödel's proof is, very roughly, the following: *Assume* that the formal system F is consistent (otherwise it proves, by elementary logic, every sentence and is trivially complete). By Gödel's self-reference lemma, one can then construct a sentence G_F that is independent of F (i.e. neither provable nor refutable in F). Thus F is incomplete. So far so good. Yet how then can one conclude that G_F is true?

Assuming that the formalized provability predicate used is normal, one can prove, even inside F, that

G_F is true if and only if F is consistent,

although neither side of the equivalence can be proved in F. Therefore, the truth of the sentence G_F is already implicitly assumed in the beginning of the proof, in the form of the assumption that F is consistent.

If it nevertheless turns out that F is inconsistent, one has to conclude that G_F is, after all, false — and provable in F, because every sentence is. The proof also goes through for a theory that is in fact inconsistent. An amusing real historical example is Quine's original version of his system ML (Quine 1940). At the end of the book, Quine presented a proof of Gödel's theorem for this system. But ML was later shown to be inconsistent by Rosser. Hence the Gödel sentence G_{ML} was actually false, whatever one's intuitions were.

In general, what evidence do we have for the belief that F is consistent? This varies enormously depending on the particular theory F in question. In the case of elementary arithmetic, the evidence for its consistency is overwhelming, and one can perhaps even say that it is known with mathematical certainty. On the contrary, this is not so for some of the strong new set theoretical systems such as ZFC + the existence of some huge cardinals. For such a system, the only evidence we have for its consistency is that it *seems* to formalize a consistent notion, and that one has not, so far, derived a contradiction from it.

In other words, we can also apply Gödel's theorem to a theory F about whose consistency we are less confident and prove the conditional: *If* F is consistent, *then* there is a true but unprovable-in-F sentence G_F (in the language of F). So what can we then say about the truth of G_F? The right conclusion is that we have exactly as much (or as little) reason to believe in the truth of G_F as we have reason to believe in the consistency of the formal system F in question. And the justification may vary considerably from theory to theory.

'Gödelian' arguments against mechanism

Gödel's theorems have also stimulated many philosophical speculations outside the philosophy of mathematics. In particular, one has repeatedly attempted to apply Gödel's theorems and demonstrate that the powers of the human mind outrun any mechanism or formal system. Such a Gödelian argument against mechanism was considered, if only in order to refute it, already by Turing in the late 1940s (see Piccinini 2003).

An unqualified anti-mechanist conclusion was drawn from the incompleteness theorems in a much read popular exposition, *Gödel's Theorem,* by Nagel and Newman (1958). Shortly afterwards, J.R. Lucas (1961) famously proclaimed that Gödel's incompleteness theorem "proves that Mechanism is false, that is, that minds cannot be explained as machines". He stated that "given any machine which is consistent and capable of doing simple arithmetic, there is a formula it is incapable of producing as being true ...but which we can see to be true". More recently, very similar claims have been put forward by Roger Penrose (1990, 1994).[7] Crispin Wright (1994, 1995) has endorsed related ideas from an intuitionistic point of view.[8] They all insist that Gödel's theorems imply, without qualifications, that the human mind infinitely surpasses the power of any finite machine. These Gödelian anti-mechanist arguments are, however, flawed.

The basic error of such an argument is actually rather simply pointed out.[9] The argument assumes that for any formalized system, or a finite machine, there exists the Gödel sentence (saying that it is not provable in that system) which is unprovable in that system, but which the human mind can see to be true. Yet Gödel's theorem has in reality the conditional form, and the alleged truth of the Gödel sentence of a system depends on the assumption of the consistency of the system. That is, all that Gödel's theorem allows us humans to prove with mathematical certainty, of an arbitrary given formalized theory F, is:

F is consistent $\Rightarrow G_F$.

The anti-mechanists argument thus also requires that the human mind can always see whether or not the formalized theory in question is consistent. However, this is highly implausible. After all, one should keep in mind that even such distinguished logicians as

[7] For detailed criticism of Penrose by experts of the field, see Boolos 1990, Davis 1990, 1993, Feferman 1995, Lindström 2001, Pudlak 1999, Shapiro 2003.

[8] For criticism, see Detlefsen 1995.

[9] This objection goes back to Putnam 1960; see also Boolos 1967.

Frege, Curry, Church, Quine, Rosser and Martin-Löf have seriously proposed mathematical theories that have later turned out to be inconsistent. As Martin Davis has put it: "Insight didn't help" (Davis 1990). Lucas, Penrose and others have certainly attempted to reply to such criticism (see e.g. Lucas 1996, Penrose 1995, 1997), and have made some further moves, but the fact remains that they have never really managed to get over the fundamental problem stated above. At best, they have changed the subject.

John Searle (1997) has joined the discussion and partly defended Penrose against his critics. It seems, though, that Searle has missed the point. He assumes that the standard criticism is based on the suggestion that the relevant knowledge might be unconsciousness. Searle argues that such a critique fails. Yet the real issue has absolutely nothing to do with awareness. Penrose's key assumption, that the algorithm or formal system must be "knowably sound", refers to the idea that one must, in addition to possessing certain axioms and rules, know that they are sound, that is, that they produce no false theorems (or at least that they are consistent). Whether this knowledge is conscious or unconscious is totally irrelevant for the main question. If our understanding would really exceed that of any possible computer, we should be able to always see whether a given formal system is sound or not. And to assume that is quite fantastic. Searle seems to uncritically accept the belief held by Penrose and others that a human being can always "see the truth" of a Gödel sentence. And, this, we have seen, is the basic fallacy in these "Gödelian" arguments for anti-mechanism.

Quite recently Storrs McCall has made an effort to provide improved Gödelian arguments against mechanism (McCall 1999, 2001). McCall admits that the standard anti-mechanist argument is problematic because the recognition of the truth of the Gödel sentence G_F depends essentially on the unproved assumption that the system F under consideration is consistent. McCall's new argument aims to show that still human beings, but not machines, can see that truth and provability part company. McCall suggests that we can argue by cases: Either F is consistent, in which case G_F is true but unprovable, or F is inconsistent, and G_F is provable but false. Whichever alternative holds, truth and provability fail to coincide. McCall concludes that human beings can see this, but a Turing machine cannot. This is, however, wrong. Any simple formal system (generated by a Turing machine) which contains elementary arithmetic can prove all these facts, too (see Raatikainen 2002). McCall (1999) has also attempted to give a more technical anti-mechanist argument. That argument is also flawed. Basically, it is based on an illegitimate conflation of Gödel sentences and Rosser sentences (see George and Velleman 2000, Tennant 2001).

Gödel on mechanism and Platonism

Interestingly, Gödel himself also presented an anti-mechanist argument although a more cautious one; it was published only in his *Collected Works*, Vol. III, in 1995. That is, in

his 1951 Gibbs lecture, Gödel drew the following disjunctive conclusion from the incompleteness theorems: "either ... the human mind (even within the realm of pure mathematics) infinitely surpasses the power of any finite machine, or else there exist absolutely unsolvable diophantine problems." Gödel speaks about this statement as a "mathematically established fact". Furthermore, Gödel concludes that philosophical implications are, under either alternative, "very decidedly opposed to materialistic philosophy". (Gödel 1951)[10]

According to Gödel, the second alternative, where there exist absolutely undecidable mathematical problems, "seems to disprove the view that mathematics is only our own creation; for the creator necessarily knows all properties of his creatures ... so this alternative seems to imply that mathematical objects and facts ... exist objectively and independently of our mental acts and decisions". Gödel was nonetheless inclined to deny the possibility of absolutely unsolvable problems, and although he did believe in mathematical Platonism, his reasons for this conviction were elsewhere, and he did not maintain that the incompleteness theorems alone establish Platonism. Thus Gödel believed in the first disjunct, that the human mind infinitely surpasses the power of any finite machine. Still, this conclusion of Gödel follows, as Gödel clearly explains, only if one denies, as does Gödel, the possibility of humanly unsolvable problems. It is not a necessary consequence of incompleteness theorems:

> However, as to subjective mathematics [PR: humanly knowable mathematics], it is not precluded that there should exist a finite rule producing all its evident axioms. However, if such a rule exists, ... we could never know with *mathematical certainty* that all propositions it produces are correct ... the assertion ... that they are all true could at most be known with empirical certainty .... there would exist absolutely unsolvable diophantine problems ..., 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. (Gödel 1951, my emphasis)

Now Gödel was, unlike the later advocated of the so-called Gödelian anti-mechanist argument, sensitive enough to admit that both mechanism and the alternative that there are humanly absolutely unsolvable problems are consistent with his incompleteness theorems. His fundamental reasons for disliking the latter alternative are much more philosophical. Gödel thought in a somewhat Kantian way that human reason would be fatally irrational if it would ask questions it could not answer. If, on the other hand, we are ready to accept a more modest view on our human capabilities, and admit that there may exist mathematical problems that are absolutely undecidable for us, this alternative causes no problems, and is indeed philosophically the easiest to accept. But does this alternative really imply, as Gödel believed, the truth of mathematical Platonism. Not necessarily. There is an option, suggested e.g. by Kreisel (1967) while commenting on Gödel's disjunctive conclusion. Kreisel writes: "I do not make the assumption that, if

[10] For more discussion on Gödel's disjunctive claim, see e.g. Shapiro 1998.

mathematical objects are our own constructions, we must be expected to be able to decide all their properties; for, except under some extravagant restrictions on what one admits as the self I do not see why one should expect so much more control over one's mental products than over one's bodily products — which are sometimes quite surprising" (Kreisel 1967). I am inclined to agree.

Actually Gödel explicitly considered this alternative in the form of following objection: "For example, we build machines and still cannot predict their behaviour in every detail". "But", Gödel continued, "this objection is very poor. For we don't create the machines out of nothing but build them out of some material" (Gödel 1951). I do not think that Gödel's reply is really convincing. He ignores the possibility of designing, for example, a computing machine in the functional level, e.g. by writing a flow chart, totally independently of the different material realizations of it. Still, the question whether a given program halts or not may be totally opaque for the programmer who has created the program. And the question is completely independent of the materials one uses to realize the program; it is a software issue independent of the hardware. In sum, I think that the alternative that there are humanly absolutely unsolvable problems does not necessarily imply Platonism.

Benacerraf, mechanism and self-knowledge

As a reaction to Lucas' argument, but before the publication of Gödel's Gibbs Lecture, Paul Benacerraf (1967) put forward more qualified conclusions that interestingly resemble some ideas of Gödel. That is, Benacerraf first argued that given any Turing machine T, either I cannot prove that T is adequate for arithmetic, or if I am a subset of T, then I cannot prove that I can prove everything T can. He concluded that it is consistent with all this that I am indeed a Turing machine, but one such that I cannot ascertain what it is. Benacerraf interprets the philosophical import of this colorfully: "If I am a Turing machine, then I am barred by my very nature from obeying Socrates' profound philosophical injunction: *Know thyself*."

Benacerraf has certainly provided a true logical fact: he shows that certain assumptions are together inconsistent. Still, it is not entirely clear what the real relevance of this is philosophically. As John Burgess has pointed out (reported in Chihara 1972), much depends on what is meant by an 'absolute proof'. If it is required that the premises of an absolute proof must be self-evident, then it is possible that I am a formalized theory, that I can discover empirically that I am the theory F, but that I cannot prove this absolutely. Kripke in turn has suggested (also reported by Chihara 1972) that the fact that I cannot discover the program does not seem to be so paradoxical when it is observed that such a discovery involves distinguishing what I can really prove (absolutely) from what I merely think I can prove. Hence it involves distinguishing such things as genuine absolute proofs from apparent proofs and genuine knowledge from mere beliefs.

Indefinite extensibility and expansion procedures

Dummett (1963) examines the intuitionist thesis that mathematical proof or construction is essentially a mental entity. He interprets this as a rejection of the idea that there can even be an isomorphism between the totality of possible proofs of statements within some mathematical theory and the proofs within any formal system. Although Dummett wants to distance himself from the psychologistic language of traditional intuitionism, he thinks that this fundamental point is entirely correct. According to Dummett (1963), Gödel's theorems shows that no formal system can ever succeed in embodying all the principles of proof of the arithmetical statements we should accept. He has influentially expressed this conclusion also by saying that the class of intuitively acceptable proofs is an *indefinitely extensible* one. By this he means, in this context, that for any formal system, once the system has been formulated, one can, by reference to it, define new properties which are not expressible in the system. Moreover, by applying induction to such new properties, one can arrive at conclusions that are not provable in the systems.

Later, Dummett (1994) has been, for good reason, more cautious. He has admitted that it does not directly follow from Gödel's theorems that the set of arithmetical truths we are capable of recognising as such cannot be recursively enumerable. According to Dummett, incompleteness theorems only rule out the possibility that in that case we can, from the specification of the set, recognise that it contains only true theorems. Dummett now submits that the standard objection to the Lucas-Penrose argument is sound and that the only conclusion we can draw is a disjunctive one (such as Gödel's above). Dummett states that the sentences which result from the indefinite iteration of this procedure "cannot all be derived within a single formal system that we can recognise as intuitively correct; ... if there is any sound formal system of arithmetic in which they can be all derived, we cannot recognise its soundness" (Dummett 1994).

Wright (1994) affirms that the sentences which result from indefinite iteration of extension procedure cannot be recursively axiomatizable. Yet, as Dummett (1994) correctly comments, all that can be concluded is that if it is recursively axiomatizable, we cannot recognise its soundness (we should add, following Gödel, recognize *with mathematical certainty*). Nevertheless, in the end, Dummett adds that "there are multifarious ways of extending an intuitively correct formal system of arithmetic". Dummett mentions, for instance, Feferman's 'autonomous progression', transfinite induction, adding truth predicate and suitable axioms, etc. He concludes after all that Gödel's theorem guarantees that we cannot encapsulate all extensions of arithmetic into a single intuitively correct formal system (Dummett 1994). It is not clear that this is necessarily true.

Feferman's work is highly relevant here. Feferman has studied various different processes of extension for formal systems. The first such approach was in terms of the autonomous

transfinite progressions of theories (Feferman 1964). Later, Feferman introduced a more general a notion of reflective closure of a system S, which used Kripke-Feferman truth theory. Feferman proposed that the reflective closure of a system S contains everything one ought to accept if one has accepted the basic notion and principles of S. Feferman also showed that the reflective closure and the earlier autonomous progression, when using PA as the initial theory, entail exactly the same arithmetical truths, i.e the theorems of the system of ramified analysis up to but not including Γ_0 (Feferman 1991).

More recently, Feferman has formulated (Feferman 1996, Feferman & Strahn 2000) a new very general notion of the 'unfolding' closure of schematically axiomatized formal systems S. It provides a uniform systematic means of expanding in an essential way both the language and axioms of such systems S. He suggests that this is even more convincing as an explication of everything that one ought to accept if one has accepted given concepts and principles. Once again, the unfolding of PA is proof-theoretically equivalent to the system of ramified analysis up to but not including Γ_0, and hence equivalent to both autonomous progression and reflective closure when applied to PA. There is thus striking stability in the end result in these very different ways of extending standard arithmetic. I think that especially together these results strongly suggest that Feferman's notions manage indeed to capture everything that is implicitly accepted when one accepts the original system.

One should note that the totality of arithmetical statements which are provable along such extension processes can nevertheless be captured by a formalized system. Furthermore, assuming that the system is consistent, one can again apply Gödel's theorem and get an arithmetical sentence which is unprovable in the system but true, although it is not — if Feferman is right — acceptable on the basis of what was implicit in the acceptance of the initial theory. But this seems to be a problem for Dummettians' idea of indefinite extensibility, especially when combined with their intuitionistic equation of truth with provability. Truth, for them, apparently cannot go beyond what is acceptable on the basis of the original concepts and principles.

Mysticism and the existence of God ?

Sometimes quite fantastic conclusions are drawn from Gödel's theorems. It has been even suggested that Gödel's theorems — if not exactly prove — at least give strong support for mysticism or the existence of God. For example, the well known popularizer of science, Paul Davies, reflecting on Gödel's results, concludes: "We are barred from ultimate knowledge, from ultimate explanation, by the very rules of reasoning that prompt us to seek such an explanation in the first place. If we wish to progress beyond, we have to embrace a different concept of 'understanding' from that of rational explanation. Possibly the mystical path is a way to such understanding. Maybe [mystical insights] provide the

only route beyond the limits to which science and philosophy can take us, the only possible path to the Ultimate." (Davies 1992).

Michael Guillen interprets the moral of Gödel's results as thus: "the only possible way of avowing an unprovable truth, mathematical or otherwise, is to accept it as an article of faith." (Guillen 1983, pp. 117-18). Juleon Schins (1997) even declares that Gödel's (and Turing's) results "firmly establish the existence of something that is unlimited and absolute, fully rational and independent of human mind". "What would be more convincing pointer to God", he asks. Antoine Suarez (1997) in turn states that, because of Gödel's theorems, we are "scientifically" led to the conclusion that it is reasonable to reckon with God.

Perhaps a person who is inclined to see evidence for God's existence everywhere can also see it in Gödel's theorems, but in themselves, these results have no such implications. Among other confusions, these interpretations seem to assume one or more misunderstandings which have already been discussed above. It is either assumed that Gödel provided an absolutely unprovable sentence, or that Gödel's theorems imply Platonism, or anti-mechanism, or both. But arguably all such conclusions are unjustified.

Bibliography

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

Boolos, George (1968) "Review of 'Minds, machines and Gödel', by J.R. Lucas, and 'God, the Devil, and Gödel', by P. Benacerraf", *Journal of Symbolic Logic* 33, 613–15.

Boolos, George (1990) "On 'seeing' the truth of Gödel sentence", *Behavioral and Brain Sciences* 13, 655–656.

Chaitin, Gregory J. (1974) "Information-theoretic limitations of formal systems", *Journal of the Association for Computing Machinery* 21, 403–24.

Chaitin, G.J. (1975) "Randomness and mathematical proof", *Scientific American* 232, 47–52.

Chaitin, Gregory J. (1987a) "Incompleteness theorem for random reals", *Advances in Applied Mathematics* 8, 119–146.

Chaitin, G.J. (1987b), *Algorithmic Information Theory*, Cambridge University Press, Cambridge, 1987.

Chihara, Charles (1972) "On alleged refutations of mechanism using Gödel's incompleteness results", *Journal of Philosophy* 69, 507–26.

Crocco, Cabriella (2003) "Gödel, Carnap, and the Fregean heritage", *Synthese* 137, 21–41.

Davies, Paul (1992), *The Mind of God,* Simon & Schuster, New York.

Davis, Martin (1990) "Is mathematical insight algorithmic?", *Behavioral and Brain Sciences* 13, 659–660

Davis, Martin (1993) "How subtle is Gödel's theorem? More on Roger Penrose", *Behavioral and Brain Sciences* 16, 611–612.

Dummett, Michael (1963) "The philosophical significance of Gödel's theorem", *Ratio* 5, 140–155. Reprinted in M. Dummett: *Truth and Other Enigmas*, Duckworth, London, 1978, 186–201.

Dummett, Michael (1994) "Reply to Wright", in Brian McGuinness and Gianluigi Oliver (eds.) *The Philosophy of Michael Dummett,* Kluwer, Dordrecht, 329–338.

Detlefsen, Michael (1995) "Wright on the non-mechanizability of intuitionist reasoning", *Philosophia Mathematica* 3, 103–118.

Feferman, Solomon (1964) "Systems of predicative analysis", *Journal of Symbolic Logic* 29, 1–30.
Feferman, Solomon (1991) "Reflecting incompleteness", *Journal of Symbolic Logic* 56, 1–49.
Feferman, Solomon (1995) "Penrose's Gödelian argument: A review of *Shadows of Mind*, by Roger Penrose", *Psyche* 2 (7).
Feferman, Solomon (1996) "Gödel's program for new axioms: why, where, how and what?", in *Gödel '96*, *Lecture Notes in Logic* 6, 3–22.
Feferman, Solomon and Thomas Strahm (2000) "The unfolding of non-finitist arithmetic", *Annals of Pure and Applied Logic* 104, 75–96.
George, Alexander and Daniel Velleman (2000) "Leveling the playing field between mind and machine: a reply to McCall", *Journal of Philosophy* 97, 456–461.
Goldfarb, Warren (1995) "Introductory note to *1953/9", in Gödel 1995, 324-334.
Warren Goldfarb and Thomas Ricketts (1992), "Carnap and the philosophy of mathematics", in David Bell and Wilhelm Vossenkuhl (eds.), *Science and Subjectivity*, Akademie Verlag, Berlin, 1992, pp. 61–78.
Gödel, Kurt (1931). "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I", *Monatshefte für Mathematik und Physik* 38, 173–98; translated in Gödel 1986*,* 144**–**195.
Gödel, Kurt (1933). "The present situation in the foundations of mathematics", in Gödel 1995, pp. 45**–**53.
Gödel, Kurt (1941) "In what sense is intuitionistic logic constructive?", in Gödel 1995, p. 189**–**200.
Gödel, Kurt (1944) "Russell's mathematical logic", in P. A. Schilpp (ed.) *The Philosophy of Bertrand Russell*, Northwestern University, Evanston, Il., 125**–**153. Reprinted in Gödel 1990, 119**–**141.
Gödel, Kurt (1951) "Some basic theorems on the foundations of mathematics and their implications" (Gibbs Lecture). In Gödel 1995, pp. 304**–**323.
Gödel, Kurt (1986). *Collected Works I. Publications 1929–1936.* ed. S. Feferman *et al.*, Oxford University Press, Oxford.
Gödel, Kurt (1990). *Collected Works II. Publications 1938–1974.* ed. S. Feferman *et al.*, Oxford University Press, Oxford.
Gödel, Kurt (1995). *Collected Works III. Unpublished Essays and Lectures*, ed. S. Feferman *et al.*, Oxford University Press, Oxford.
Guillen, Michael (1983) *Bridges to Infinity*. Tarcher, Los Angeles.
Hellman, Geoffrey (1981) "How to Gödel a Frege-Russell: Gödel's incompleteness theorems and logicism", *Nous* 15, 451–468
Henkin, Leo (1962) "Are mathematics and logic identical?", *Science* 138, 788–794.
Kreisel, Georg (1967) "Mathematical logic: what has is done for the philosophy of mathematics?", in Ralph Schoenman (ed.) *Bertrand Russell. Philosopher of the Century.* George Allen & Unwin, London, 201–272.
Lindström, Per (2001) "Penrose's new argument", *Journal of Philosophical Logic* 30, 241–250.
Lucas, J. R. (1962) "Minds, machines, and Gödel", *Philosophy* 36, 112-137.
Lucas, J. R. (1996) "Minds, machines, and Gödel: A retrospect", in P.J.R. Millican and A. Clark (eds.) *Machines and Thought. The Legacy of Alan Turing,* Vol. 1, Oxford University Press, Oxford, 103–124.
McCall, Storrs (1999) "Can a Turing machine know that the Gödel sentence is true?", *Journal of Philosophy* 96, 525–532.
McCall, Storrs (2001) "On 'seeing' the truth of Gödel sentence", *Facta Philosophica* 3, 25–29.
Musgrave, Alan (1977) "Logicism revisited", *British Journal for the Philosophy of Science* 28, 99–127.
Nagel, Ernest and James R. Newman (1958). *Gödel's Proof*, New York University Press, New York.
Penrose, Roger (1989) *The Emperor"s New Mind: Concerning Computers, Minds, and the Laws of Physics***,** Oxford University Press, New York.
Penrose, Roger (1994) *Shadows of the Mind: A Search for the Missing Science of Consciousness***,** Oxford University Press, New York.

Penrose, Roger (1995) "Beyond the doubting of a shadow: A reply to commentaries of *Shadows of the Mind*", *Psyche* Vol 2.
Penrose, Roger (1997) "On understanding understanding", *Inernational Studies in the Philosophy of Science* 11, 7-20.
Piccinini, Gualtiero (2003) "Alan Turing and the mathematical objection", *Minds and Machines* 13, 23–48.
Pudlak, Pavel (1999) "A note on applicability of the incompleteness theorem to human mind", *Annals of Pure and Applied Logic* 96, 335–342.
Putnam, Hilary (1960) "Minds and machines", in S. Hook (ed.), *Dimensions of Mind*, New York University Press, New York, 1960. Reprinted in H. Putnam: *Mind, Language, and Reality.* Philosophical Papers, Vol 2. Cambridge University Press, Cambridge, 1975, 325–341.
Putnam, Hilary (1967) "Mathematics without foundations", *Journal of Philosophy* 44, 5–22. Reprinted in H. Putnam: *Mathematics, Matter and Method. Philosophical Papers Vol 1.* Cambridge University Press, Cambridge, 1975, 43–59.
Putnam, Hilary (1975) "What is mathematical truth?", *Historia Mathematica* 2, 529–545. Reprinted in H. Putnam: *Mathematics, Matter and Method. Philosophical Papers Vol 1.* Cambridge University Press, Cambridge, 1975, 60–78.
Quine, W. V. (1940) *Mathematical Logic*, Harvard University Press, Cambridge, MA.
Quine, W. V. and J. S. Ullian (1978) *The Web of Belief.* 2nd ed, Random House, New York.
Raatikainen, Panu (1998) "On interpreting Chaitin's incompleteness theorem", *Journal of Philosophical Logic* 27, 269–586.
Raatikainen, Panu (2000) "Algorithmic information theory and udecidability", *Synthese* 123, 217–225.
Raatikainen, Panu (2001) "Review of *The Unknowble* and *Exploring Randomness*", *Notices of the American Mathematical Society,* Volume 48, Number 9, 992–996.
Raatikainen, Panu (2002) "McCall's Gödelian argument is invalid", *Facta Philosophica* Vol. 4, No 1, 167–169.
Raatikainen, Panu (2003a) "Hilbert's program revisited", *Synthese* 137, 157–177.
Raatikainen, Panu (2003b) "Some strongly undecidable natural arithmetical problems, with an application to intuitionistic theories", *Journal of Symbolic Logic* 68, 262–266.
Raatikainen, Panu (2004) "Conceptions of truth in intuitionism", *History and Philosophy of Logic* (forthcoming)
Reinhardt, William (1985) "Absolute versions of incompleteness theorems", *Noûs* 19, 317–46.
Ricketts, Thomas (1995) "Carnap's principle of tolerance, empiricism, and conventionalism", in Peter Clark & Bob Hale (eds.), *Reading Putnam*, Blackwell, Cambridge, 1995, 176-200.
Rodríguez-Consuegra, Francisco (1993) "Russell, Gödel and logicism", in J. Czermak (ed.), *Philosophy of mathematics*. Hölder-Pichler-Tempsky, Vienna, 1993, 233–42. Reprinted in A. Irvine (ed.), *Bertrand Russell: Critical Assessments*. Routledge, London and New York, 1998, vol. 2: *Logic and mathematics,* 320–29.
Rosser, Barkley (1936) "Extensions of some theorems of Gödel and Church", *Journal of Symbolic Logic* 1, 87–91.
John Searle (1997) "Roger Penrose, Kurt Gödel, and the Cytoskeletons", in J. Searle: *Mystery of Consciousness*, New York Review Books, New York, 55–93.
Shapiro, Stewart (1998) "Incompleteness, mechanism, and optimism", *Bulletin of Symbolic Logic* 4, 273–302.
Shapiro, Stewart (2003) "Mechanism, truth and Penrose's new argument", *Journal of Philosophical Logic* 32, 19–42.
Schins, Juleon M. (1997) "Mathematics: a pointer to an independent reality", in A. Driessen and A. Suarez (eds.), *Mathematical Undecidability, Quantum Nonlocality and the Question of the Existence of God*, Kluwer, Dordrecht, 49–56.

Sternfeld, Robert (1976) "The logistic thesis", in Mathias Schirn (ed.) *Studien zu Frege/Studies on Frege I*, Frommann-Holzboog, Stuttgart-Bad Cannstatt, 139–160.
Suarez, Antoine (1997) "The limits of mathematical reasoning: in arithmetic there will always be unsolved solvable problems", in A. Driessen and A. Suarez (eds.), *Mathematical Undecidability, Quantum Nonlocality and the Question of the Existence of God*, Kluwer, Dordrecht, 41–48.
Tennant, Neil (2001) "On Turing machines knowing their own Gödel-sentences", *Philosophia Mathematica* Vol. 9, 72–79
Wright, Crispin (1994) "About 'The philosophical significance of Gödel's theorem': some issues", in Brian McGuinness and Gianluigi Oliver (eds.) *The Philosophy of Michael Dummett*, Kulwer, Dordrecht, 167–202.
Wright, Crispin (1995) 'Intuitionists are not (Turing) machines', *Philosophia Mathematica* 3, 86-102.