Fabrice Pataut: “In Defence of a Princess Margaret Premise,” Introductory chapter to *Truth, Objects, Infinity – New Perspectives on the Philosophy of Paul Benacerraf*, Fabrice Pataut, editor. Springer International Publishing, Logic, Epistemology, and the Unity of Science, vol. 28, Switzerland, 2016, pp. xvii-xxxviii.

ISBN 978-3-319-45978-3	ISSN 2214-9775
ISBN 978-3-319-45980-6 (eBook)	ISSN 2214-9783 (electronic)
DOI 10.1007/978-3-319-45980-6	

Library of Congress Control Number : 2016950889

IN DEFENCE OF A PRINCESS MARGARET PREMISE

Fabrice Pataut

1. Introductory remarks

In their introduction to the volume *Benacerraf and his Critics*, Adam Morton and Stephen Stich remark that "[t]wo bits of methodology will stand out clearly in anyone who has talked philosophy with Paul Benacerraf": (i) "[i]n philosophy you never prove anything; you just show its price," and (ii) "[f]ormal arguments yield philosophical conclusions only with the help of hidden philosophical premises" (*Morton and Stich 1996b*: 5).

The first bit will come to some as a disappointment and to others as a welcome display of suitable modesty. Frustration notwithstanding, the second bit suggests that, should one stick to modesty, one might after all prove something just in case one discloses the hidden premises of one's chosen argument and pleads convincingly in their favor on independent grounds.

I would like to argue that a philosophical premise may be uncovered — of the kind that Benacerraf has dubbed "Princess Margaret Premise" (PMP)[1] — that helps us reach a philosophical conclusion to be drawn from an argument having a metamathematical result as one of its other premises, viz. Gödel's first incompleteness theorem. To be somewhat more (im)precise, something philosophical may be inferred from Gödel's THEOREM VI (*Gödel [1931] 1986*: [187] 173) (and its proof) with respect to the vexed question whether truth may transcend recognizability in principle by us, humans.

Obviously, this first approximation of the question in such unblushingly Dummettian terms of transcendence, truth, recognizability, and the in principle *vs.* effective distinction must be found wanting. It is, of course, unspecific to a fault to speak in such general terms of *the* grand philosophical problem of the relation between truh and the recognition of truth, but it isn't for that matter either dreadfully vague or offensively inexact. The argument I shall propose, provided it is indeed one, has three virtues: (i) it shows with a non-negligible degree of accuracy what is the price one has to pay for the philosophical conclusion I believe may be secured; (ii) it is genuinely philosophical (as opposed to genuinely metamathematical) and does indeed lead to a conclusion, so that something in philosophy may be proved after all; (iii) thanks to (i)-(ii), the argument enables us to come up with a more refined version of the problem I just started with without thereby loosing sight of its very general scope. I shall explain why it is important not to loose sight of such a scope in the concluding remarks, i.e. why, although powerful formal tools play a crucial role in the obtaining of a sober and strictly philosophical conclusion, the larger philosophical picture, for the benefit of which "a lot of delicate informal interpretation" (*Morton and Stich loc. cit.*) has to be put to good use, *must* in the end matter to us (see *infra* Section 5).

The only rather feeble apology I am able to offer at this point with regard to the unreliability of the introductory formulation of the question to be answered is that the forthcoming argument reveals "unexpected premises and consequences" (*Morton and Stich loc. cit.*) at play in the controversy about the independence of truth from our ability to recognize that truth obtains when it does, so that, if the argument does indeed go through, we shall at least end up with an improved formulation of the philosophical puzzle we started with.[2]

2.1 Correctness and truth

Let me start with Gödel's result proper and with remarks — some by Gödel, some by others — pertaining to it that will play a role in the forthcoming argument.

Gödel certainly thought it worthwhile to remind his readers that his proof of the first incompleteness theorem was *constructive*. He pointed to the fact that the result had been obtained "in an intuitionistically unobjectionable manner" (*Gödel [1931] 1986*: [189] 177) and offered as a warrant for this claim that "all existential statements [*Existentialbehauptungen*] occurring in the proof [were] based upon THEOREM V [i.e the theorem immediately preceding the first incompleteness theorem] which, as is easily seen, is unobjectionable from the intuitionistic point of view" (*Gödel loc. cit.*: note 45a). In Kleene's terminology, THEOREM V states that every primitive recursive relation is numeralwise expressible in *P*, where *P* is the system obtained from

Whitehead and Russell's *Principia Mathematica*, without the ramification of the types, taking the natural numbers as the lowest type and adding their usual Peano axioms (*Kleene 1986*: 129). When expressed formally, without reference to any particular interpretation of the formulas of *P*, and in Gödel's own terminology which favors the indirect talk of Gödel numbers and of concepts applying to these numbers rather than a direct talk of the formal objects (i.e. the formulas and the variables), THEOREM V claims that:

> For every recursive relation $R(x_1,\ldots, x_n)$ there exists an n-place RELATION SIGN r (with the FREE VARIABLES u_1, $u_2,\ldots$, u_n) such that for all n-tuples of numbers $(x_1,\ldots, x_n)$ we have
>
> $$R(x_1,\ldots, x_n) \rightarrow \text{Bew}\ [Sb(\text{r}^{u_1\ \cdots\ u_n}_{Z(x_1)\ldots\ Z(x_n)})],$$
>
> $$\overline{R}(x_1,\ldots, x_n) \rightarrow \text{Bew}\ [\text{Neg}(Sb(\text{r}^{u_1\ \cdots\ u_n}_{Z(x_1)\ldots\ Z(x_n)}))].$$
>
> *Gödel ([1931] 1986*: [186] 171

Gödel sketches an outline of the proof and notes on this occasion that THEOREM V is itself "of course, [...] a consequence of the fact that in the case of a recursive relation *R* it can, for every *n*-tuple of numbers, be decided *on the basis of the axioms of the system P* whether the relation *R* obtains or not" (*Gödel op. cit.*: [186n39] 171n39). This, it must be noted, may also be decided by means of procedures that remain unobjectionable from the intuitionistic standpoint.

Gödel's true and undecidable formula, the existence of which is proved constructively by the first incompleteness theorem (THEOREM VI) may seem at first sight to offer a counterexample to the claim that truth may not transcend recognition by us, either in principle or effectively, for the proof establishes the existence of a formula which does have both properties, viz. that of truth *and* that of undecidability. Gödel's diagonal argument does indeed provide a true statement which is nevertheless omitted by the relevant algorithm.[3]

Two aspects of the situation passed on to us by Gödel's proof somewhat complicate the matter. First, there are followers of Wittgenstein's *Remarks on the Foundations of Mathematics* like Shanker who think that it is incoherent and indeed downright nonsensical to claim that a statement or formula is (or may be, for that matter) both true and undecidable. Shanker points out that, if a complete manifestation of our recognition of the truth of the Gödel formula were possible, the semantic formulation of the theorem would thereby be defective: it would turn the connection between a mathematical statement and its proof into a purely external matter (*Shanker 1990*: 221ff). This strongly suggests that the first incompleteness theorem should be formulated in syntactical fashion, without reference or commitment to truth, as stating that every formal system S, if consistent, and when elementary number theory is taken as its domain, contains a formula $\mathcal{A}$ expressing a proposition A of elementary number theory such that neither $\mathcal{A}$ nor its negation $\neg\mathcal{A}$, expressing $\neg A$, is provable in S.

So, to begin with: may we or may we not claim that Gödel's undecidable formula is true *simpliciter*, or true *tout court*, as distinct from either recognizably true or, on the contrary, true beyond recognition, either effective or in principle, by us, humans? When giving an informal sketch of the main idea of the proof in the first section of his 1931 paper, Gödel says that if the proposition [$R(q);q$] were provable, it would also be correct [*richtig*] and that, in that case, $\overline{\text{Bew}}$ [$R(q);q$] would hold [*würde gelten*], “which contradicts the assumption” (*Gödel op. cit.*: [175] 149.[4] If, on the other hand, the negation of [$R(q);q$] were provable, Bew [$R(q);q$] would hold. Then both [$R(q);q$] and its negation would be provable, “which again is impossible.” He then concludes this introductory section by saying that “[f]rom the remark that [$R(q);q$] says about itself that it is not provable, it follows at once that [$R(q);q$] is [correct] [*richtig ist*], for [$R(q);q$] *is* indeed unprovable (being undecidable). Thus the proposition that is undecidable *in the system PM* still was decided by metamathematical considerations” (*Gödel op. cit.*: [176] 151).

Gödel does not use the German *wahr* in this instance. Jean van Heijenoort resorts to the English equivalent of that German word in his translation (which I have departed from here on purpose) and Kleene, in his presentation, also claims that the formula $\mathcal{A}$ is “unprovable, hence *true* [emphasis mine]” (*Kleene 1986*: 128).[5] Is it sensible, then, to claim that truth forces its way into the Gödelian picture — which after all is entirely in terms of a proposition being *correct* and in terms of a claim to the effect that x is not a provable formula (or, better, in terms of a claim to the effect that a natural number q belongs to a class K of natural numbers, with K defined in terms of non provability) *holding* — *only* because of van Heijenoort's translation, and that Kleene's presentation is, likewise, flawed, or at least anomalous in this respect?[6]

It is not, even though there is *no* notion of "correctness" or of "holding" to be *mis*translated in THEOREM VI itself:

> For every ω-consistent recursive class k of FORMULAS there are recursive CLASS SIGNS r such that neither v Gen r nor Neg (v Gen r) belongs to Flg(k) (where v is the FREE VARIABLE of r).

Although Gödel's formulations do not involve a direct or explicit claim to the effect that the undecidable formula is true, or true without any further proviso or qualification, but only a claim to the effect that, for all x, x isn't the Gödel number of a proof of it, it can hardly be maintained that the formula which is undecidable modulo the consistency and ω-consistency of P, and which states that it is neither provable nor refutable in the system, may *not* be a truth bearer (and thus, may *not* be true *simpliciter* and, a fortiori true *and* undecidable).

As far as the informal presentation is concerned, the undecidable formula *truly* says of itself that it is not provable for it *is* indeed not provable. Is the situation any different when, instead of referring to the undecidable formula by means of its metamathematical description [$R(q);q$], we refer to it by means of its Gödel number once we have determined the number q, i.e. by the expression "17 Gen r" ("x Gen y" denoting the 15th number theoretic function proven to be (primitive) recursive)? Undeniably, Gödel concludes his proof of the first incompleteness theorem by saying that "17 Gen r is therefore undecidable on the basis of k, which proves THEOREM VI" (*Gödel op. cit.*: [189] 177) and *not* by saying that “17 Gen *r* is therefore true and undecidable on the basis of *k*, which proves THEOREM VI.” The undecidable formula nevertheless *truly* claims that 17 Gen *r* is not *k*-PROVABLE and that Neg (17 Gen *r*) is, likewise, not *k*-PROVABLE.

Of course, Gödel remarks that “the purpose of carrying out the […] proof with full precision […] is, among other things, to replace the [assumption that every provable formula is *true* [[my emphasis]] in the interpretation considered] by a purely formal and much weaker one” (*Gödel op. cit.* : [176] 151). Kleene might be right to explain that we assumed that only true formulas are provable in *S* to the extent that this assures us that “the formulas have clear meanings” (*Kleene loc. cit.*), but this is so provided that only true formulas are provable *no matter how one refers to them*. The important point here is not about about meaning, or limpidity of whatever must be grasped. What counts is that it must not matter in this respect that 17 Gen *r* is a sentential formula whose undecidability has to be stated as being about this particular SENTENTIAL FORMULA. Kleene is certainly right to remind us at this point that the informal concept of truth was not “commonly accepted as a definite mathematical notion, especially for systems like [*Principia Mathematica*] or Zermelo-Fraenkel set theory” (*Kleene loc. cit.*). It nevertheless remains that whether we are dealing with a metamathematical description of the undecidable proposition or with the undecidable proposition itself (see *Gödel op. cit.*: [175n13] 149n13 for the distinction), we are indeed in a situation where contradicting (simple) consistency would yield the *falsity* of [*R*(*q*); *q*] (and, likewise, the falsity of 17 Gen *r*), and where contradicting ω-consistency would similarly yield the *falsity* of its negation (and, likewise, the falsity of the negation of 17 Gen *r*), as well as the falsity of an assertion to the effect that the negation of these formulas are, respectively, provable and *k*-PROVABLE.

The assertion of its own unprovability and irrefutability qualifies $\mathcal{A}$ for the status of truth bearer and the price one must pay for the jettisoning of consistency and ω-consistency is indeed, by parity, that of its falsity. Bivalence, here, is not the issue. The issue is whether the purely formal and much weaker assumpion that has replaced the informal and stronger assumption that every provable formula is true has disposed of our problem, or dissolved it into thin air, and my point here is that it hasn't.[7]

2.2 *The logical constants*

Another set of remarks that turn out to be relevant for the forthcoming argument concerns Gödel's conception that "intuitionistic logic, as far as the calculus of propositions and of quantification is concerned, turns out to be rather a renaming and reinterpretation than a radical change in classical logic" (*Gödel [1941] 1995*: [3] 190). Gödel defended this view after the publication of the undecidability results, first in a paper given at Karl Menger's colloquium in Vienna in 1932 (*Gödel [1933] 1986*), then in a lecture delivered at Yale in April 1941, from which the last quote is taken. He then discussed the view with respect to the issue of the use of abstract intuitionistic proofs in the explanation of the intuitionistic logical constants in the *Dialectica* paper from

1958 (*Gödel [1958] 1990; see also Gödel [1972a] 1990*). What is of interest to us here is that Gödel, building on results by Glivenko, showed that the classical propositional calculus is a subsystem of the intuitionistic propositional calculus and that every valid classical formula also holds in Heyting's propositional calculus provided that we translate the classical "notions" or "terms" (Gödel's words), or "operators" (Kleene's word), i.e. the classical constants, into the intuitionistic ones (*Glivenko 1929*; *Gödel [1933] 1986*).

In particular, Gödel defends the somewhat surprising claim that the law of excluded middle is intuitionistically acceptable. In classical propositional logic, ~ being the classical negation sign, the formula $p \mathrm{v} {\sim} p$ is a tautology. According to Gödel, although intuitionists reject this law for *their* notion of disjunction, one may nevertheless define *another* notion of disjunction in terms of the other primitive logical constants of *their* calculus so that, ¬ being the intuitionistic negation sign, $p \mathrm{v} \neg p$ is *also* a tautology. It is sufficient, Gödel claims, to define, quite trivially, pvq as $\neg(\neg p . \neg q)$; $\mathrm{pv}\neg\mathrm{p}$ may then be translated, just as trivially, into $\neg(\neg p . \neg\neg p)$ so that the law of excluded middle turns out to be a special case of the law of contradiction, which is, of course, intuitionistically valid (*Gödel [1941] 1995*: [2] 190).[8]

This noticeably overlooks the fact that intuitionists will want to reject classical deduction rules such as double negation elimination and classical tautologies such as excluded middle precisely because they contend that such rules and tautologies would allow us to draw *illegitimate* inferences (from judgments to judgments) on the basis of relations of *alleged* logical consequence holding between antecedent propositions and consequent propositions. What they will object to, while still holding on to the law of contradiction, is the very idea that either *p* or its negation is true, or holds, whether or not we could either be able to decide the truth-value of *p* or that of its negation. What is at stake here is the contention that the truth of *p*, or that of its negation, is independent and indeed cut off from all links with our ability to decide the matter one way or the other. Moreover, intuitionists require that the assertion of a negated proposition be justified by a *reductio ad absurdum* of the supposition that we could obtain a proof of the proposition, or of the supposition that the means of obtaining such a proof are, at least in principle, at our disposal. This strongly suggests that neither disjunction nor negation may be translated in the way suggested.

In yet other words, the following translation manual:

CLASSICAL LOGICAL CONSTANTS

1	2	3	4
$\sim p$	$p \rightarrow q$	$p \vee q$	$p \& q$

INTUITIONISTIC LOGICAL CONSTANTS

1	2	3	4
$\neg p$	$\neg(p.\neg q)$	$\neg(\neg p.\neg q)$	$p.q$

overlooks the fact that, although no special symbol for intuitionistic disjunction is thereby introduced, the claim to the effect that $p \vee \neg p$ is valid is now tantamount to the claim that we have a constructive proof of *p*v~*p*, or at least the means of obtaining one. But the claim that we have a constructive proof of *p*v~*p* is acceptable only provided that we have either a proof of *p* or a proof of ~*p* and this is definitely *not* what we have with classical disjunction. It seems, as a matter of fact, that what we have now with the translation manual is rather a claim to the effect that, independently of our knowledge, and perhaps indeed unbeknownst to us, either there is a proof of *p* or there is a proof of its negation. Or perhaps what we have is a claim to the effect that either we have a proof of *p* or we have a proof that we shall never have a proof of its negation, i.e. a proof of the double negation of *p*. But, clearly, the second term of this disjunction must be rejected by the intuitionist for a proof of the double negation of *p doesn't* amount, intuitionistically, to a proof of *p* (see endote 8).

In other words, we are surreptitiously appealing to an objective realm of proofs which is disconnected and indeed "cut off from all links with the reflecting subject," to borrow Bernays' apt phrase in his description of platonism in the philosophy of mathematics (*Bernays [1935] 1983*: 259 ; see also *Bernays op. cit.*: 267). In doing so, we are appealing to a notion of proof that is obviously *not* faithful to the constuctive standpoint.[9] If this is the case, the Glivenko-Gödel translation manual leaves both the classical logician and the intuitionistic logician unsatisfied, precisely because, most vividly in the case of excluded middle, both will judge that the meaning imposed upon disjunction and negation by way of the translation manual is arbitrary.

Although it may seem at first blush that the validity of excluded middle does not depend upon any peculiarity in the interpretation of disjunction that intuitionists would object to on the ground that the translation manual would propose a definition of disjunction "in terms of *their* [emphasis mine] other primitive logical symbols" (*Gödel [1941] 1995*: [1-2] 190), i.e. in terms of intuitionistic conjunction and negation, we are indeed in the situation that Dummett describes thus:

> The failure of the law of excluded middle is often explained by the different meaning of intuitionistic disjunction: a proof of *A*v*B* is a proof either of *A* or of *B*, and hence a claim to have proved *A*v¬*A* amounts to a claim either to have proved *A* or to have proved ¬*A*. Such an explanation of the matter is correct as far as it goes, but it will naturally leave a platonist with the feeling that the meaning imposed upon v is arbitrary: on any view on which either *A* or ¬*A* must be true, irrespective of whether we can prove it, to repudiate that sense of v in which we can assert *A* v ¬*A* a priori is to deny ourselves the means of expessing what we are able to apprehend.
>
> *Dummett 1977*: 18

There is in particular, at this fundamental level of primitive logical laws, something that seems to undermine the translation claim inherited from Glivenko and that Dummett doesn't underline in his objection, namely a disagreement about the logical form that a proprer *reductio* should take and which directly concerns negation. It must be remarked, and indeed stressed in this instance, that a classical *reductio* is *not* equivalent to an intuitionistic one and that each yields a particular form of negation, so that ~ may not, after all, be translated into ¬. The intuitionist's rationale for the rejection of *both* pv~p and pv¬p *indiscriminately* under the proposed translation scheme is that neither formulation of excluded middle amounts to the claim that we have either obtained a proof of p or a proof of its negation, or that we are in a position to obtain either. So the distinction between the allegedly objectionable pv~p and the quite acceptable pv¬p modulo the Glivenko translation turns out to be invidious: it points quite unfairly to a difference that, as a matter of fact, does not exist at all.

Let me now take stock of what has been established in section 2. The first point is that $\mathcal{A}$ is a truth bearer ; the second is that the validity of a logical law shouldn't depend on any peculiarity in the assignment of a meaning to the main constant occurring in a logically valid statement or formula that prevails as a law. The first point matters because it licenses us to give a semantic formulation of Gödel's first incompleteness theorem. The second matters because, as Gödel insists, the purpose of carrying out of the proof of THEOREM VI with full precision is to replace the assumption that every provable formula is true in the interpretation by a *weaker* one. There is no doubt that the precision is indeed provided by the proof. The point here is that there shouldn't be any arbitrariness in our determining whether or not the notion of truth involved in the claim that $\mathcal{A}$ truly asserts its own unprovability and irrefutability is itself contrained by provability or, on the contrary, unconstrained so that the truth of that claim could after all remain beyond recognizability by us, humans. As the following section should make clear, the notion of truth involved in the recognition of the truth of the proposition that asserts its own unprovability with no faulty circularity isn't arbitrarily construed. If it were, we would have begged the question. With these points in mind, let me now turn to the semantic formulation.

3.1 Gödel's theorem (I): A semantic formulation

The main outline of the semantic formulation of Gödel's proof may be given in the following way.[10] The proof exhibits an elementary formula which is finitary in Hilbert's sense and proven to be both unprovable and irrefutable in P, as the result of the following steps:

(*1*) A formula $\mathcal{A}$ is constructed by diagonalization, which asserts its own unprovability in P.

(*2a*) The consistency of P being taken for granted, it is proven that $\mathcal{A}$ is unprovable in P.

(*2b*) Since $\mathcal{A}$ asserts its own unprovability in P, $\mathcal{A}$ is true.

(*3*) The ω-consistency of P being taken for granted, it is proven that $\mathcal{A}$ is irrefutable in P.

(*4*) $\mathcal{A}$ is proven to be undecidable in P.

(*5*) $\mathcal{A}$ is true and undecidable in P.

Our warrant for (*5*) is that Gödel's formula $\mathcal{A}$ is both recognized to be true (at step (*2b*)) and proven to be undecidable (at step (*4*)) since it is both proven to be unprovable (at step (*2a*)) and proven to be irrefutable (at step (*3*)). Our question from Section 1 is whether truth may transcend its recognizability by us, humans. We have already remarked how large that question is; not just large but truly inordinate. A sense of balance or proportion might perhaps be restored if we anwer two questions that are related to it according to the description of the doctrine Dummett has dubbed “realism.”

The first question is whether our understanding of the meaning of $\mathcal{A}$ amounts to a knowledge of its truth conditions.[11] In particular, are we able to manifest, make plain or display that we do possess that knowledge? We must answer this question in the positive since the truth of $\mathcal{A}$ is *recognized* as obtaining at step (*2b*).[12] We do have a means at our disposal to find out that the truth conditions of $\mathcal{A}$ are satisfied and are able to make that knowledge manifest by proving the unprovability of $\mathcal{A}$ under the assumption that *P* is consistent (step (*2a*)) and by concluding that $\mathcal{A}$ is true (step *2b*)). Our answer is positive because, as Gödel points out (in relation to the Epimenides paradox) :

> [we] can construct propositions which make statements about themselves, and, in fact, these are arithmetic propositions which involve only recursively defined functions, and therefore are undoubtedly meaningful statements. It is even possible, for any metamathematical property *f* which can be expressed in the system, to construct a proposition which says of itself that it has this property.
>
> *Gödel [1934] 1986*: [21] 362-363

It is therefore possible, for any predicate *F* of the language L_P of *P* expressing in *P* a given metamathematical property, to construct by diagonalization a formula $\mathcal{A}$ of L_P which asserts of itself that it possesses that property. If we note the Gödel number of that formula with the symbol "$<\mathcal{A}>$," then, for every predicate F of L_P, there exists a formula such that $\mathcal{A} \Leftrightarrow F(<\mathcal{A}>)$.

Let us choose as a metamathematical property the property of non-provability in P, expressed in P by the predicate "non-Pr_P." We may then construct a formula $\mathcal{A}$ which asserts its own unprovability in P, such that $\mathcal{A} \Leftrightarrow$ non-$\mathrm{Pr}_P(<\mathcal{A}>)$.

Once that first step is accomplished, we may proceed to step (*2a*) and distinguish the following sub-steps leading to (*2b*).

If $\mathcal{A}$ were provable in P, then:

(*2a1*) $\mathrm{Pr}_P(<\mathcal{A}>)$ would be true in P and, therefore, provable in P, and

(*2a2*) non-$\mathrm{Pr}_P(<\mathcal{A}>)$ would be provable in P, since $\mathcal{A}$ and non-$\mathrm{Pr}_P(<\mathcal{A}>)$ are logically equivalent.

(*2a3*) P would therefore be inconsistent.

(*2a4*) Under the assumption that P is consistent, $\mathcal{A}$ is therefore unprovable in P.

Gödel notes in this respect that:

> Contrary to appearances, such a proposition [which says of itself that it is not provable] involves no faulty circularity, for initially it [only] asserts that a certain well-defined formula (namely the qth formula in the lexicographic order by a certain substitution) is unprovable. Only subsequently (and so to speak by chance [*gewissermaßen zufällig*]) does it turn out that this formula is precisely the one by which the proposition itself was expressed.
>
> *Gödel [1931] 1986*: [176n15] 151n15

We may then directly proceed to step (*2b*): since $\mathcal{A} \Leftrightarrow$ non-$\mathrm{Pr}_P(<\mathcal{A}>)$, $\mathcal{A}$ is true. It is thus clear that the question whether or not we know the truth conditions of $\mathcal{A}$ may — and indeed must — be answered in the positive by the time we reach (*2b*), for we can make it perfectly plain, by proceeding from step (*2a1*) to step (*2b*) that we know indeed that these truth conditions are satisfied. So it is possible for us, humans, to manifest (to borrow once again from Dummett's terminology) our knowledge of the truth conditions of a formula proven to be unprovable in a formal system obtained from Whitehead and Russell's *Principia*, without the ramification of the types, that takes the natural numbers as the lowest type and incorporates the usual Peano axioms.

As far as truth conditionality is concerned, we are not in a situation where it would be appropriate to eschew the twin notions of truth and truth conditions altogether. Since arithmetic propositions that involve only recursively defined functions are "undoubtedly meaningful statements" (*Gödel [1934] 1986*: [21] 362), there is no reason to jettison the truth conditionality principle, as applied to $\mathcal{A}$. The meaning of $\mathcal{A}$ is indeed constituted by its truth conditions and our knowledge of that meaning amounts to a knowledge of these conditions. There is indeed something in virtue of which $\mathcal{A}$ is true, i.e. something in virtue of which its truth conditions are fulfilled, namely the proof that proceeds from (*1*) to (*2b*).

3.2 Gödel's theorem (II): Two gaps

The second question related to the description of the doctrine Dummett has dubbed "realism" is whether we are in a case where truth transcends, one way or another, recognition by us, humans: either recognition in principle (in which case *theoretical or ideal recognizability* is at stake), or outright effective (in which case *actual or feasible recognition* is at stake). This, of course, takes us back to the vexed question we started with in section 1, but we are now in a somewhat more comfortable position for we may after all get a purchase on that controversy. Gödel's proof is unambiguous in this respect : the elementary formula proven to be undecidable in *P* given the consistency and ω-consistency of *P* is true in a sense that may not offend either a constructivist or Dummett's antirealist. As noted at the beginning of section 2.1, Gödel thought it worthwhile to remark that his proof was unobjectioanble from the intuitionistic standpoint. In particular, the proof doesn't allow one to conclude that the truth conditions of $\mathcal{A}$ transcend its justification conditions ; it shows something quite different, namely that:

> […] our capacities for justification go beyond what is strictly speaking provable in a formal system: there exists, for each sufficiently rich formal system [i.e. such that the property "provable in the system" is expressible in the system], undecidable elementary statements that we nevertheless have cogent reasons to hold as true.
>
> *Dubucs 1991*: 57[13]

In other words, Gödel's first incompleteness theorem doesn't show, either directly or indirectly — i.e. either with or without the help of a Princess Margaret Premise — that the extension of the predicate "true" is larger than the extension of the predicates "recognizable (in principle) as true" and "(effectively) recognized as true." The first gap would indeed be unacceptable from a constructivist or antirealist standpoint, the second from a strict finitist one. What the proof of the theorem shows is that the extensions of the last two predicates are larger than the extension of the predicate "provable in P," which is quite another matter. What we have been able to acknowledge so far is that the truth conditions of $\mathcal{A}$ are transcendent with respect to its provability in P, but this offers nothing in the way of an admission of some 'absolute' notion of recognition-transcendent truth, of some supreme notion, as it were, of *truth beyond all possible justification*, or even of the possibility thereof.

Two conclusions may thus be drawn.[14] The first is that there is no elementary formula whose truth could be undetectable in a formal system if we assume that system to be consistent (which we do). The most we are allowed to say is that there are elementary formulas whose truth remains algorithmically undetectable given the consistency of the system, and that amounts to a quite different claim. In the case in point, no algorithmic procedure may help us to conclude that $\mathcal{A}$ is true, but its truth is nevertheless acknowledgeable by us by means of a *reductio ad absurdum* of the supposition that it is provable in P, given that P is consistent. Unless we decree that the unavailability of an algorithmic procedure for deciding the truth-value of a formula is a criterion for the undetectability of its truth, there are no undetectable truths, or truths beyond all possible recognition in a formal system if that system is consistent. So the question is: Should we order the decree? Step (*2b*), from the previous section, strongly suggests that we may not even argue for this position (let alone decree that we may benefit from such a criterion). We must on the contrary distinguish the case of algorithmic undecidability from that of undetectability of truth-value *simpliciter* (or, better, in the case at hand, of undetectability of truth-value by any *non*-algorithmic means) when discussing the vexed question we started with.

The second conclusion to be drawn is thus that we must take into account a finer-grained distinction than the one we have been pondering over so far, i.e. one that contrasts:

(A) The gap between what is true in the standard model for arithmetic and what is recognizable as true on the basis of cogent reasons

with

(B) The gap between what is recognizable as true on the basis of cogent reasons and what is algorithmically recognizable as true in the standard model for arithmetic.

The first gap is filled by the proof of the unprovability of $\mathcal{A}$ and, a fortiori, by the (complete) proof of its undecidability. The second may *not* be filled, just because of the very same proof.

The question now is: What does this tell us about the vexed question we started with?

4.1 The missing Princess Margaret Premise: Benacerraf's assessment

When discussing the case of Gödel's incompleteness results, and of the first result in particular, Benacerraf drops more than a gentle hint at what the Princess Margaret Premise is, and indeed should be, with respect to the anti-mechanistic conclusion that we are not machines, especially when that conclusion is based on the "libertarian arguments to the effect that our abilities transcend those of any machine, outstrip in truth-power any formal system, i.e. do not constitue or cannot be adequately represented as an i.e. set of sentences etc." (*Benacerraf 1996b*: 42-43). The proprer PMP one would have to add to the formal result in order to get the desired conclusion is that:

> There is something human mathematicians can do that no machine can do — for any (theorem proving) machine, find its Gödel number and, given its Gödel number, prove its Gödel sentence (something it manifestly cannot do).
>
> *Benacerraf op. cit.*: 31

Although the finer-grained distinction does allow us to conclude that there is something human mathematicians can do that no machine can do, it does *not* thereby support the much stronger view that *we are not machines*. For the sake of simplicity, let us call the claim that we must take into account the distinction between (A) and (B), along with the remarks about gap-filling (and the impossibility of gap-filling) offered at the very end of section 3.2, the "two gaps thesis." We are *not* in a situation where we could avail ourselves to a specific instance of an argument of the form:

(C) *(Metamathematical result, PMP)* ⊢ *Conclusion*

with :

(C*) (THEOREM VI, Two gaps thesis) ⊢ *We are not machines* in the role of the desired specific instance.

If we were arguing in this way, we would indeed be deriving an unwarranted philosophical conclusion. Benacerraf knows this, of course, and remarks in this respect that:

> The [first incompleteness] theorem heralds what its name suggests: an incompleteness in formalized arithmetic. However hard we may squeeze, we can not extract from it the thought that, although we once believed we had a concept of arithmetic truth, now that we see that the sentences true in arithmetic cannot be exactly those corralled by any plausible formal "proof" procedure (if bivalence is also to be preserved), we must concede that we did not. The incompleteness that Gödel demonstrated, the incompleteness of the First Incompleteness Theorem, was shown to exist in the calculus, not in our conception. Not that one couldn't be lurking in our conception as well — of course one could — but that just hasn't been shown; [...] to go that extra mile requires a Princess Margaret Premise and a separate argument to support it.
>
> *Benacerraf op. cit.*: 42

The two gaps thesis supports a much weaker view. It is therefore crucial, in order to identify that view correctly, to avoid either crediting the human mind with cognitive capacities it doesn't have, or denying that a Turing machine associated with the relevant formal system of arithmetic may perform tasks that it is, after all, clearly able to perform. In particular, it might be objected that although we're able to provide a justification for $\mathcal{A}$, given that the formula correctly expresses its own unprovability, the availability of such a justification through a non-mechanizable step doesn't thereby establish that we're not a Turing machine, but only that we're not a Turing machine *associated with* P. After all, since P is taken to be consistent and proves $Cons(P)\rightarrow\mathcal{A}$, the system $P^* = P\cup\{Cons(P)\}$ also proves $\mathcal{A}$. A machine that would enumerate the arithmetical theorems of P^* would indeed be able to generate $\mathcal{A}$. We wouldn't, therefore, be in a case where some arithmetical truth has been omitted and where the human mind would thereby be cognitively "superior" to the machine associated with P^*.

The anti-mechanist or libertarian might object at this point that she is not arguing that there exists an elementary formula of arithmetic whose truth a human mind is able to justify but that no consistent Turing machine will ever engender (or that such a machine will necessarily omit). She might wish to claim that her argument establishes the falsity of the general claim that Turing machines may, qua consistent, enumerate all the arithmetical truths for which (non-mechanical) human minds are able to find, albeit non mechanically, a justification. She would then be proposing, in order to stand her ground, a case-by-case refutation of *each particular instance* of the general mechanistic claim. This would imply that the mechanism involved in the libertarian claim to be rejected and the libertarian argument to be refuted, amounts to the rather weak proposal that the mind is a machine, *but only given that it is established that, for each particular machine we may care to consider when assessing the mind-machine metaphysical identity thesis, the mind isn't* that *particular machine*. Some ω-inconsistency would indeed be involved in such a conception of the non-mechanistic mind.[15]

Notice that Benacerraf, in *Benacerraf 1967*, seems to argue in favor of such a thesis: he infers from Gödel's incompleteness results that we are not Turing machines, *or at least that, if we were, we wouldn't be able to determine our own instruction tables*. This, of course, is a crucial proviso for it seems indeed to be a way of saying that the mind is a machine, albeit with the rather damaging caveat that it may be established, for each particular machine, that the mind isn't *that* machine, just because the mind cannot acknowledge what its own Turing instructions are and is therefore clearly deficient in terms of self- or introspective knowledge.

We are left, then, with a position that draws as a conclusion from Gödel's first incompleteness result some indulgent form of libertarianism based on a weak notion of mechanism; a notion so week indeed that it cannot do justice to the idea at the heart of libertarianism that our cognitive abilities transcend those of any Turing machine and, in particular, that such machines are unable to enumerate all the formulas whose truth a human mind or agent can acknowledge or recognize insofar as the mind or agent targets *the standard model for arithmetic*.

There must be another way to do justice to Benacerraf's important remark that Gödel's incompleteness should not be located in our *conception* of arithmetical truth and should safely remain where it belongs, i.e. in the *calculus* (or in the family of *calculi*) we've managed to devise. What is the view, then, which the two gaps thesis may support, so that one isn't either "brandishing" the metamathematical result as the authority for some purely philosophical yet unwarranted conclusion (*Benacerraf op. cit.*: 43), or defending one that involves ω-inconsistency?

4.2 In defence of the proper Princess Margaret Premise

It looks like I've driven myself into a tight corner. I've claimed in section 4.1 that there is something we can do that Turing machines cannot do, i.e. fill the first gap, and that this distinction supports a philosophical view which is both weaker than the strictly anti-mechanistic or libertarian view, and distinct from the one Benacerraf defends — or at least seems to defend — in *Benacerraf 1967*. Whatever its details might turn out to be, the philosophical conclusion we may draw must be true to Benacerraf's brief that the incompleteness is in the calculus and not in our conception of arithmetical truth. If the PMP is indeed one, the philosophical conclusion it yields must be consistent with the view that our conception of arithmetical truth, or of what makes a statement of arithmetic true when it is true, should be the very notion we had before Gödel proved his incompleteness result, so that the "unformalized practice of mathematics" (*Benacerraf op. cit.*:12) must escape unscathed from the Gödelian result. The point I wish to make with respect to the two gaps thesis being the proper PMP we need is that, contrary to what Benacerraf contends, something may *not* remain unscathed, namely the "implications […] regarding the very nature of […] ourselves as [the] practitioners [of mathematics]" (*Benacerraf op. cit.*: 14). Although we must reject the claim that our former conception of arithmetical truth was defective, despite the undeniable fact, established by

Gödel's result, that "the sentences true in arithmetic cannot be exactly those corralled by any formal 'proof' procedure (if bivalence is also to be preserved)" (*Benacerraf op. cit.*: 42), we still must conclude that there is *a human cognitive capacity that transcends those of any Turing machine*. It is the conception of "ourselves," or of our minds, or of the scope and nature of our own cognitive capacities that must be amended so that the incompleteness is, strictly speaking, a property of the mathematical formalism, and nothing more.

Benacerraf claims that the argument that purports to show that we are

> subject to the same limitations that have been proved to hold of the formal languages and systems that we study in metamathematics (or, in other cases, free from them) […], if it is to be at all probative […], must include a convincing demonstration of the relevant isomorphism (or lack thereof) between our own powers and the relevant features of the systems.
>
> (*Benacerraf op. cit.*: 12)

The lack of isomorphism is manifested by our capacity to fill the gap between what is true in the standard model for arithmetic and what is recognizably true on the basis of cogent reasons, and this is precisely what allows us to conclude that we're not machines, albeit neither in the strong metaphysical sense that we're not strictly speaking identical to any Turing machine, nor in the weaker sense that we're not any particular machine we may care to consider, given our ignorance, in each and every particular putative case, of our own instruction tables. The crucial point here is that Gödel's incompleteness proof, by showing that the first gap is filled, *thereby shows that the second may not be filled*. So our PMP relies crucially on the claim that we have the cognitive capacity to target *the* standard model for arithmetic, for it is this very capacity which is responsible for our acknowledgement of the truth of $\mathcal{A}$, i.e., the capacity that so troubles Shanker (see, supra, section 2.1).

To go back to the first Benacerrafian methodological point that Morton and Stich remind us of (see *Morton and Stich 1996b*: 5) and that I've mentioned at the very beginning of this paper in the first paragraph of section 1, it would be strange indeed, perhaps even mystifying, if nothing philosophical could be "proven" or argued in this regard, so that nothing more than the price of the philosophical point about ourselves, or our minds, or our cognitive capacitiy to target the standard model of arithmetic, would be known. After all, if we've shown the price of what must be proven, we know the price, and if we know the price, the only thing that could prevent us from coming by the conclusion is some unfortunate lack of funds. But how could that be? If we have determined which argument would yield a philosophical view when appended to an established metamathematical result, then we do, *de facto*, have that argument in favor of the purely philosophical conclusion, although not necessarily an independent one for each of its premises.[16]

Benacerraf remarks, of course, that the metamathematical result *alone* yields a *weaker* purely philosophical conclusion (*Benacerraf op. cit*: 43), so that if (C), then:

(C') *Metamathematical result* ⊢ *(PMP → Conclusion)* in which case

> (C'*) (THEOREM VI) $\vdash$ (Two gaps thesis $\rightarrow$ *We are not machines*

would be the specific desired libertarian instance.

I've argued here, though, in favor of a different and somewhat weaker claim. We do need, of course, independent grounds for that claim, grounds which would give it a bona fide proper content: we do need to say more on the nature of these "cogent" reasons. I won't be arguing for such independent grounds here. What may be stressed, though, is that since $\mathcal{A}$ is a genuine truth bearer and the notion of truth involved in the claim that $\mathcal{A}$ truly asserts its own unprovability isn't either guilty of faulty circularity or arbitrarily construed, Gödel's proof, together with the two gaps thesis yields a claim to the effect that we are free of one limitation that has been proved to hold for *P*, where *P* is the system obtained from *Principia Mathematica*, without the ramification of the types, taking the natural numbers as the lowest type and adding their usual Peano axioms.

We cannot garantee any public display of our non-mechanical knowledge that the truth conditions of $\mathcal{A}$ obtain, given that *P* is consistent, over and above what the argument sketched in section 3.1 imparts with steps (*1*)-(*5*) (and steps (*2a1*)-(*2a4*)). This, obviously, is a defect for which the semantic formulation argued for in section 2.1 is fully responsible. Shanker, or anyone convinced by his skepticism, may rightly object that since the reference to the standard model may not be eliminated from our argument, and since no recursively axiomatizable class of formulas allows us to give a proper definition of such a standard model, the argument is wanting unless it may be shown, once again on independent grounds, that our grasp of the standard model need *not* rely on the existence of a language fit to describe it, say second order Peano arithmetic, or some other.[17]

What I would like to stress here in way of a conclusion is that, perhaps more than truth proper, the crucial philosophical dimension of the semantic reading of Gödel's first incompleteness theorem is bivalence, as the remarks of section 2.2 about the meaning of the logical constants in relation to the validity of logical laws, and the quote, in this section, to the effect that arithmetic truth doesn't quite match arithmetic provability if bivalence is to be secured (*Benacerraf 1996b*: 42) clearly indicate. I shall now turn to this question.

5. Concluding remarks about the larger picture

I promised in section 1 to explain why it is important not to loose sight of the scope of the general question we started with in spite of the fact that it might appear too general, unspecific and, for this very reason, nothing less than ill-defined. There is no doubt that the question whether truth may go beyond recognizability in principle is a crucial philosophical question. The problem, rather, if the gist of the argument that unfolds from the beginning of section 3.1 to the end of section 4.2 stands up to critical examination, is whether we should care about that larger picture at all in the specific case of the Gödelian metamathematical result we've taken into consideration.

The large question has gained momentum and even currency in great part because of Dummett's persistent claim that it lies at the core of the realism *vs*. antirealism debate, conceived as a debate whose genuine content is semantic in nature (hence the reference to those "Dummettian terms of transcendence, truth, recognizability and the in principle *vs*. effective distinction" in the opening paragraphs of section 1). Part of the advantage of being at this center is that it discloses the non-metaphorical content of the time-honored metaphysical issue of the existence and, should existence be granted, of the independence of objects of some given kind, or sort, or class ; typically, abstracta such as numbers, but also, say, colors or values. Since semantics is a stake, the question of bivalence is crucial; not just bivalence, but that of semantic principles in general. More specifically, what is at stake is the question of their relation to logical laws. How is, say, bivalence related to excluded middle, or stability to double negation elimination? One idea is that the semantic principles justify the logical laws. So the question is: should be accept the semantic principles?

Benacerraf rejects the idea that since

> (bivalent) truth outruns formal provability, for any consistent formal system of Arithmetic, the concept (of bivalent truth) is inherently flawed and must be replaced with a concept of truth in arithmetic that is more closely tailored to our ability as provers. This, of course, then splinters into countless possibilities, depending on how bounded we believe our "proving" abilities to be.
>
> *Benacerraf op. cit.*: 30

The splintering might well be curtailed so that, after all, only two main possibilities remain, or at least two kinds thereof, depending on whether one rests content with in principle formal provability, or wishes to insist that effective formal provability, garanteed to be humanly feasible, must be secured in order for the principle of bivalence to be acceptable.

Given what has been said before about the implications of Gödel's result given some proprer PMP, this conception of the issue at stake strongly suggests that we are facing a dilemma: either we must conclude that we are somehow free of the shackles built in the mechanistic view because we are endowed with super-mechanical powers, or truth must be constrained by some epistemic notion such as provability (either in principle, which would satisfy Dummett's antirealist, or effective, along strict finitist lines). In other words, it is as if we should choose between two conclusions to be derived, provided one may be derived at all with the aid of some bona fide PMP: either the adoption of a strong metaphysical libertarian thesis, or the replacement of bivalent truth conditions by conditions such that it must be garanteed, by the very nature of the case, that we, humans, are able to recognize that they obtain when they do (or at least that we are able to put ourselves in the position to activate the appropriate recognitional capacities).

It is, I suspect, because of such a dilemma (either libertarianism or the death of bivalence) that Benacerraf takes Putnam to task for locating the incompleteness in our conception on the ground of an endorsement of a "heavily 'epistemic' notion of truth" (*Benacerraf op. cit.*: 44).

Benacerraf's rationale for rejecting Putnam's negative conclusion regarding bivalence is that epistemic considerations, as applied to truth — and hence to the issue of the unrestricted appeal to the semantic principle —, rather than offering reasons to be restrictive about what we are able to understand "point the opposite way." His justification for this position is that he

> [takes] "epistemic" considerations to include ones of meaning, but without accepting as axiomatic that these are ineluctably entwined with a semantics of "assertibility conditions," as opposed to a "truth-conditional" semantics, or even some other, as yet undreamt-of, kind.
>
> *Benacerraf op. cit.*: 44.

It might seem odd, once epistemic considerations have been taken to include those of meaning, especially with respect to the issue of grasp or understanding, to claim that the latter are divorced from considerations pertaining to assertibility conditions. It is odd, of course, not because one would be an antirealist in Dummett's prefered sense, but because our cognitive limitations as "provers," i.e. those that play a key role in the PMP, come in a twosome, along with our *non*-mechanistic ability to fill the gap between what is true in the standard model and what is recognizable as true on the basis of cogent reasons (provided we're able to target the standard model and that the targeting need not rely on a language fit to give a bona fide description of it). These limitations come hand in hand with, as it were, a positive aptitude or capacity. The reason why we are not Turing machines after all is that being such a machine requires that there be a finite collection of instructions, each instruction calling for atomic operations to be performed under certain given conditions. The recognition of such a cognitive limitation or failure, is therefore crucial, in Benacerraf's 1967 moderate anti-mechanistic conclusion. This is even more so in the argument I have sketched: there is a bound to our purely mechanical proving abilities, and a corresponding release, as it were, of our non mechanistically characterizable bona fide recognitional capacities with respect to truth.

Putnam, quoted by Benacerraf, claims that since nothing epistemic may help us "explain the truth-value of [...] undecidable statements, precisely because they *are* undecidable," i.e. no matter how hard we constrain the notion of arithmetic truth, we should refrain from attaching any "metaphysical weight to the principle of bivalence," a principle which would have us believe that these statements are either true or false although "their truth-value cannot be decided on the basis of the axioms we presently accept" (*Benacerraf op. cit.*: 44).

What I have been urging here, the acceptance of axioms notwithstanding, and the crucial matter of the targeting of the standard model pending, is to leave bivalence alone and to draw from a modest PMP an equally modest conclusion to the effect that there is a least one capacity that the anti-libertarian or hard-nosed computationalist is unable to account for.

Notes

1. Benacerraf reports the following parable of the Cohens and Princess Margaret (*Benacerraf 1996b*: 9-10). The Cohens want the right spouse for their son. When they accept somewhat reluctantly the goy Princess Margaret as the right girl for Abie on the basis of the staggering advantages their future grandchildren would benefit from (being heirs to the throne of England, etc.), only *half* the shatchen's job is done. Abie still has to marry Princess Margaret. Without the Cohen's acceptance of the marriage broker's final offer, no result could be obtained, but more is needed nevertheless for the job to be rounded off. The Cohen's reluctant acceptance after the turning down of so many proposals is what Benacerraf mischievously dubs "the easy part"; what is needed now is the extra move which will bring the broker's efforts to its expected conclusion, i.e. a marriage settlement.

In the same way, once you've got your metamathematical result and the first half of the job is done, you still need some extra premise — the so-called "Princess Margaret Premise" —, along with an independent argument in its favor, to get the analogue of the settlement, i.e. the desired philosophical conclusion that you may not obtain directly from the metamathematical result.

2. An important part of what follows draws from *Pataut 1998*. That paper did not address Benacerraf's points about the PMP one must fall back on in order to draw philosophical conclusions from Gödel's first incompleteness theorem, although it did address the truth *vs.* recognition of truth issue. My purpose here is to address as explicitely as possible the points Benacerraf makes with respect to this opposition, especially the point about the demonstrated incompleteness being shown to exist *in the calculus* and not *in our conception* (see *Benacerraf op. cit.*: section 4.2, pp. 42-44). It is essential to avoid both the mathematical fallacy of deducing philosophical conclusions directly from formal results of a metamathematical kind, and the philosophical blunder of mistaking some irrelevant extra premise for the genuine Princess Margaret one.

3. Boolos has proposed a non-constructive proof in *Boolos 1989*. His proof, just like Gödel's, establishes the existence of an undecidable statement of arithmetic, but, unlike Gödel's, it does not provide an effective procedure for producing it. Let a correct algorithm M be an algorithm which may not list a false statement of arithmetic. A truth omitted by M is a true statement of arithmetic not listed by M. Boolos's proof establishes the existence of such a statement, but the statement is recognized as true classically and not constructively.

4. *Richtig* may also be rendered as "right," and *würde gelten* as "would be valid," or "would obtain."

5. $\mathcal{A}$ states that every natural number x is not the Gödel number of a proof, in S, of a formula that turns out to be $\mathcal{A}$.

6. Kleene notes that Gödel approved van Heijenoort's translation of his 1931 paper for the then forthcoming volume van Heijenoort was editing

(*van Heijenoort 1967*), and that the translation was accomodated in many places to Gödel's own wishes (*Kleene 1986*: 141).

[7] Gödel notes that "the true reason for the incompleteness inherent in all formal systems of mathematics is that the formation of ever higher types can be continued into the transfinite, while in any formal system at most denumerably many of them are available" (*Gödel [1931] 1986*: [191] 181, note 48a).

Kleene's reading of this remark is that Gödel implicity defends the view that the adjunction of higher types "permits one to define the notion of truth for that system, then to show that all its provable sentences are true and hence to decide the sentence shown in Theorem VI to be undecidable in the system" (*Kleene 1986*: 135). In that case, then, the sentence (or formula) whose undecidability has been decided thanks to the adjunction of higher types, is indeed a bona fide true *and* undecidable sentence (or formula), as suggested towards the beginning of this section but, in that instance, without the recourse to types.

[8] $\neg (\neg p . \neg\neg p)$ is a case of the law of contradiction just in case double negation may be eliminated, something an intuitionist will *not* grant.

[9] This, of course, must be judged against Gödel's complaint that, it is "doubtful whether the intuitionists have really remained faithful to their constructive standpoint in setting up their logic" and, worse, that "the notion of an intuitionistically correct proof or constructive proof lacks the desirable precision" to the point that "it furnishes itself a counterexample against its own admissibility, insofar as it is doubtful whether a proof utilizing this notion of a constructive proof is constructive or not" (*Gödel [1941] 1995*: [3-4] 190).

[10] It is a much too simple-minded formulation that doesn't do justice to essential features of Gödel's proof such as the Gödel numbering of the formal objects, the notion of numeralwise expressibibiliy and the constructively defined notion of the class of primitive recursive functions. Its one and only purpose is to focus on the relation between arithmetical truth and formal proof procedures in formalized arithmetic.

[11] For the idea that $\mathcal{A}$ has a meaning, or is meaningful, see Section 2.1 above and Kleene's remark in *Kleene 1986*: 128, already quoted in that section.

[12] The claim should be qualified. At steps (*2a*)-(*2b*), we manifest our knowledge that *if P* is consistent, then $\mathcal{A}$ is unprovable in *P* and therefore true. There remains the further problem of knowing how we could know that *P* is consistent and make that knowledge manifest. I have focused here on the consequent of the conditional, but it is of course established by Gödel's second incompleteness theorem (Theorem XI in *Gödel [1931] 1986*) that a proof of the consistency of *P* may not be obtained in *P*. What we have here, therefore, strictly speaking, is only a *partial* manifestation of our knowledge of the truth conditions of $\mathcal{A}$.

[13] My translation from the French.

14. The claim should be qualified. The conclusions follow from the first part of Gödel's proof and are grounded on the steps of its semantic formulation up to (*2b*). I have not taken into consideration the proof of the unprovability of $\mathcal{A}$ (given the ω-consistency of *P*). Note that the very same conclusions would a fortiori be justified were the complete proof taken into consideration.

15. See *Dubucs 1992*: 75-76, to which I am very much indebted for these remarks.

16. See *Benacerraf 1967* and *Benacerraf 1996b*: 54, endnote 20, for a stern dismissal of arguments *without* PMPs, i.e., without "a significant injection of philosophical serum."

17. See *Dummett [1963] 1978* for a discussion of the manifestation or public warrant of our private but still commonly shared grasp of bona fide mathematical objects and formal proofs, in regard to Gödel's incompleteness result.

References

Benacerraf (Paul), *1967*, "God, the Devil and Gödel," *The Monist*, vol. 51 (January), pp. 9-32.

- *1996b*, "What Mathematical Truth Could Not Be – I," in *Benacerraf and his Critics*, A. Morton and S. P. Stich, eds., Blackwell Publishers, Oxford and Cambridge, Mass., pp. 9-59.

Bernays (Paul), *[1935] 1983*, "On platonism in mathematics," English translation by C. D. Parsons, in *Philosophy of mathematics - Selected readings*, P. Benacerraf and H. Putnam, eds., Cambridge UP, Cambridge, 2nd ed., pp. 258-271.

Boolos (George), *1989*, "A new proof of the Gödel incompleteness theorem," *Notices of the American Mathematical Society*, Vol. 36 (4), pp. 388-390.

Dubucs (Jacques), *1992*, "Arguments gödéliens contre la psychologie computationnelle," *Travaux de logique n° 7 (juin 1992) : Kurt Gödel, Actes du colloque, Neuchâtel, 13 et 14 juin 1991*, Denis Miéville, éd., Centre de Recherches Sémiologiques, Université de Neuchâtel, pp. 73-89.

Dummett (Michael A. E., Sir), *[1963] 1978*, "The philosophical significance of Gödel's theorem," in *Truth and Other Enigmas*, 1st edition, Duckworth, London, pp. 186-201.

- [with the assistance of Roberto Minio)], *1977*, *Elements of intuitionism*, Oxford Logic Guides n°2, 1st edition, Clarendon Press, Oxford.

Glivenko (Valerii I.) (Гливéнко, Валéрий Ивáнович), *1929*, "Sur quelques points de la logique de M. Brouwer," *Académie royale de Belgique, Bulletin de la classe des sciences*, Vol. 15 (5), pp. 183-188.

Gödel (Kurt), *[1931] 1986*, "Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I" / "On formally undecidable propositions of *Principia mathematica* and related systems I," in *Collected Works*, Vol. I: *Publications 1929-1936*, S. Feferman, Ed.-in-chief, Oxford UP, Oxford, English translation by J. van Heijenoort, pp. [173-198] 144-195.

- *[1933] 1986*, "Zur intuitionistischen Arithmetik und Zahlentheorie" / "On intuitionistic arithmetic and number theory," in *Collected Works*, Vol. I: *Publications 1929-1936*, S. Feferman, Ed.-in-chief, Oxford UP, Oxford, English translation by S. Bauer-Mengelberg and J. van Heijenoort, pp. [34-38] 287-295.

- *[1934] 1986*, "On undecidable propositions of formal mathematical systems [with the Postscriptum (3 June 1964)]," in *Collected Works*, Vol. I: *Publications 1929-1936*, S. Feferman, Ed.-in-chief, Oxford UP, Oxford, pp. [1-27] 346-371.

- *[1941] 1995*, "In what sense is intuitionistic logic constructive?", in *Collected Works*, Vol. III: *Unpublished essays and lectures*, S. Feferman, Ed.-in-chief, Oxford UP, Oxford, pp. [1-30] 189-200.

- *[1958] 1990*, "Über eine bisher noch nicht benütze Erweiterung des finiten Standpunktes" / "On a hitherto unutilized extension of the finitary standpoint," in *Collected Works*, Vol. II: *Publications 1938-1974*, S. Feferman, Ed.-in-chief, Oxford UP, Oxford, English translation by S. Bauer-Mengelberg and J. van Heijenoort, pp. [280-287] 241-251.

- *[1972a] 1990*, "On an extension of finitary mathematics which has not yet been used," in *Collected Works*, Vol. II: *Publications 1938-1974*, S. Feferman, Ed.-in-chief, Oxford UP, Oxford, English translation by L. F. Boron, revised by A. S. Troelstra, pp. 271-280.

Heijenoort (Jean van), ed., *1967*, *From Frege to Gödel: a source book in mathematical logic, 1879-1931*, Harvard UP, Cambridge, Mass.

Kleene (Stephen C.), *1986*, "Introductory note to *1930b, 1931* and *1932b*," in Kurt Gödel: *Collected Works*, Vol. I: *Publications 1929-1936*, S. Feferman, ed.-in-Chief, Oxford UP, Oxford, pp. 126-141.

Morton (Adam) and Stich (Stephen P.), eds., *1996a*, *Benacerraf and his Critics*, Blackwell Publishers, Oxford, and Cambridge, Mass.

Morton (Adam) and Stich (Stephen P.), *1996b*, "Introduction," in A. Morton and S. P. Stich eds., pp. 1-5.

Pataut (Fabrice), *1998*, "Incompleteness, Constructivism and Truth," *Logic and Logical Philosophy*, Vol. 6, R. Leszko and H. Wansing, Guest Editors, pp. 63-76.

Shanker (Stuart G.), *1990*, *Gödel's Theorem in focus*, Routledge, London.