POLSKA AKADEMIA NAUK, INSTYTUT MATEMATYCZNY

DISSERTATIONES MATHEMATICAE
(ROZPRAWY MATEMATYCZNE)

CXVIII

G. KREISEL and G. TAKEUTI

Formally self-referential propositions for cut free classical analysis and related systems

WARSZAWA 1974
PAŃSTWOWE WYDAWNICTWO NAUKOWE

PRINTED IN POLAND

WROCŁAWSKA DRUKARNIA NAUKOWA

CONTENTS

INTRODUCTION

Typical examples of self-referential propositions for a formal system $\mathscr{F}$ assert:

I am — or: I am not — formally derivable in $\mathscr{F}$ (from some assumption A). Or, a little more precisely, they assert:

One, or each, of my names in $\mathscr{F}$ is — resp. is not — derivable in $\mathscr{F}$ etc.

Such propositions and their variants have been applied with excellent results, above all by Gödel [5] and Löb [18], to establish the underivability of certain instances of so-called *reflection* or soundness principles for $\mathscr{F}$ in the sense of [16], namely (variants of) the schema which expresses, for some subclass of formulae A of $\mathscr{F}$:

If A is derivable in $\mathscr{F}$, then (the proposition expressed by) A is true.

Below we shall write $\vdash_{\mathscr{F}} A$ for $\exists x \mathrm{Prov}_{\mathscr{F}}(x, \ulcorner A \urcorner)$, where $\mathrm{Prov}_{\mathscr{F}}$ and $\ulcorner A \urcorner$ are the *canonical representations* (in the sense of [11]; p. 154, 3.222) of the derivability predicate of $\mathscr{F}$ and of the syntactic object A resp.

For the applications [1] intended by Gödel to the 'usual' systems, which in fact satisfy the conditions listed in Section 2 of the Appendix (or, by Löb, in [18]) there is no need for a close study of self referential propositions, in particular of the exact syntactic form of formulae which *express* their own (un)derivability. Specifically, let A_G and A_H be any two formulae for which

$$A_G \leftrightarrow \neg \vdash_{\mathscr{F}} A_G \quad \text{and} \quad A_H \leftrightarrow \vdash_{\mathscr{F}} A_H$$

are derivable in $\mathscr{F}$. Then so are

$$A_G \leftrightarrow \mathrm{Con}_{\mathscr{F}} \quad \text{and} \quad A_H \text{ itself},$$

where $\mathrm{Con}_{\mathscr{F}}$ expresses (canonically) the consistency of $\mathscr{F}$. In other words all such A_G (or 'Gödel sentences for $\mathscr{F}$') are demonstrably equivalent, namely to $\mathrm{Con}_{\mathscr{F}}$ and so are all such (Henkin sentences; cf. [6]) A_H since they are demonstrably true. (Gödel sentences are of course also true but

---

[1] A review of this work and especially of its significance for Hilbert's programme, one of the principal concerns of [5], is given in Part III. Since [5] appeared 40 years ago, it would be sad if there had not been enough progress to justify some reformulations of [5].

not demonstrable in $\mathscr{F}$.) These equivalences are mildly unexpected since the (?) propositions asserted when the two authors of this article say

I am Japanese

are demonstrably not equivalent, and the propositions expressed by two distinct Gödel or Henkin sentences are.

We do not propose to go here into the significance of those equivalences. But we do answer the question

How general are these equivalences?

by establishing that the restriction to the usual systems is essential. We shall be mainly concerned with Gödel and Henkin sentences of *cut free analysis* $\mathscr{CFA}$ described in Part I (in contrast to 'usual' or 'full' analysis $\mathscr{UA}$), but also with so-called Rosser variants of these systems which have been used in [11], p. 154, 3.221, and [8], p. 299, to analyze the conditions for the validity of Gödel's second incompleteness theorem concerning the formal underivability of $\mathrm{Con}_{\mathscr{F}}$ in $\mathscr{F}$ (2). A Gödel sentence of the Rosser variant of a system $\mathscr{F}$ will also be called *Rosser sentence of* $\mathscr{F}$.

The use of $\mathscr{CFA}$ in this connection recommends itself for two reasons. Firstly, $\mathscr{CFA}$ unlike Rosser variants is not a system manufactured for the express purpose of technical counterexamples, but is the fruit of one of the most striking logical discoveries since [5], the analysis of reasoning by 'cut free' or 'normal' derivations (though the explicit analyses of the logical interest of those discoveries differ remarkably for different 'cut free' systems; cf. III.2, that is, Section 2 of Part III, below). Secondly, $\mathscr{CFA}$ may be expected to present novelties for self-referential propositions since — as the name 'cut free' suggests — $\mathscr{CFA}$ is not demonstrably closed under cut and this fact is related to equivalence of Gödel sentences A_G and A'_G, as follows:

Suppose $\vdash_{\mathscr{F}}(A_G \to A'_G)$ but *not* $\vdash_{\mathscr{F}}(A'_G \to A_G)$ and hence, since A_G and A'_G are Gödel sentences for $\mathscr{F}$, *not* $\vdash_{\mathscr{F}}[(\neg \vdash_{\mathscr{F}} A'_G) \to (\neg \vdash_{\mathscr{F}} A_G)]$. Then $\mathscr{F}$ is not demonstrably closed under cut because

$$\vdash_{\mathscr{F}}(A_G \to A'_G) \text{ by hypothesis, but } \textit{not } \vdash_{\mathscr{F}}[(\vdash_{\mathscr{F}} A_G) \to (\vdash_{\mathscr{F}} A'_G)].$$

(Incidentally, this remark provides yet another use of Gödel sentences for incompleteness proofs.)

---

(2) In [20] there is no mention of a new formal system (whereas, for us, Rosser's own formula is simply a Gödel sentence for a less symmetric form of what we here call Rosser variant; for a systematic discussion see A.3). Throughout we formulate results for *canonical* representations of formal systems presented by formal rules as data, and not by manufactured 'representations' of the set of theorems generated by the rules — in contrast to e.g. [2].

More generally, if $\mathscr{F}$ is not demonstrably closed under cut, then — in sharp contrast to such usual systems as $\mathscr{U}\mathscr{A}$ — *formal equivalence* between formulae A and B, that is

$$\vdash_{\mathscr{F}}(A \leftrightarrow B)$$

does not generally imply their *deductive equivalence*, that is

$$\text{derivability in } \mathscr{F} \text{ of } [(\vdash_{\mathscr{F}} A) \leftrightarrow (\vdash_{\mathscr{F}} B)]$$

(even though for all the system $\mathscr{F}$ considered in this paper,

$$(\vdash_{\mathscr{F}} A) \leftrightarrow (\vdash_{\mathscr{F}} B) \text{ will be } \textit{true} \text{ whenever } \vdash_{\mathscr{F}}(A \leftrightarrow B)$$

holds). The reader will recall that the proofs of equivalence between Gödel sentences, resp. between Henkin sentences for the usual systems use heavily the passage from formal to deductive equivalence. In contrast we have the following *principal results* for $\mathscr{C}\mathscr{F}\mathscr{A}$.

The *Gödel sentences* for $\mathscr{C}\mathscr{F}\mathscr{A}$ are precisely those for $\mathscr{U}\mathscr{A}$. They are, therefore, demonstrably in $\mathscr{C}\mathscr{F}\mathscr{A}$, formally and deductively equivalent to $\mathrm{Con}_{\mathscr{C}\mathscr{F}\mathscr{A}}$ which, in turn, is equivalent to $\mathrm{Con}_{\mathscr{U}\mathscr{A}}$. In contrast, the Gödel sentences of the Rosser variant $\mathscr{C}\mathscr{F}\mathscr{A}_R$ (defined in I.1.8), that is, the *Rosser sentences for* $\mathscr{C}\mathscr{F}\mathscr{A}$, are not all equivalent; some are, formally and deductively, equivalent to $\mathrm{Con}_{\mathscr{C}\mathscr{F}\mathscr{A}}$, others — like *all* Rosser sentences for $\mathscr{U}\mathscr{A}$ — strictly weaker than $\mathrm{Con}_{\mathscr{C}\mathscr{F}\mathscr{A}}$. Specifically, *literal* Rosser sentences for $\mathscr{C}\mathscr{F}\mathscr{A}$ are equivalent to $\mathrm{Con}_{\mathscr{C}\mathscr{F}\mathscr{A}}$ where, generally for any system $\mathscr{F}$, a sentence A is called a *literal Gödel sentence* for $\mathscr{F}$ if

> A is literally of the form $\neg\vdash_{\mathscr{F}} A$, that is $\neg\exists x \mathrm{Prov}_{\mathscr{F}}(x, \ulcorner A \urcorner)$ not merely formally equivalent to $\neg\vdash_{\mathscr{F}} A$ ([3]).

The *Henkin sentences* for $\mathscr{C}\mathscr{F}\mathscr{A}$ — in contrast to those for $\mathscr{U}\mathscr{A}$ — are not all equivalent; in particular, $0 = 0$ is a Henkin sentence for $\mathscr{C}\mathscr{F}\mathscr{A}$, as it is for $\mathscr{U}\mathscr{A}$, but also $\neg 0 = 0$ (which is not a Henkin sentence for $\mathscr{U}\mathscr{A}$). More interestingly — and in sharp contrast to $\mathscr{U}\mathscr{A}$ where, by [18], *all* Henkin sentences are derivable — all *literal Henkin sentences for* $\mathscr{C}\mathscr{F}\mathscr{A}$ *are actually refutable in* $\mathscr{C}\mathscr{F}\mathscr{A}$. For a formal system in which not even all literal Henkin sentences are equivalent, see the end of A.2 (that is, of Section 2 in the Appendix). This supersedes [10].

The results above also show that the obvious proposals for extending Löb's theorem [18] to $\mathscr{C}\mathscr{F}\mathscr{A}$ are false; cf. I.2.4 (that is, Section 2.4 of Part I). Thus provided $\mathscr{C}\mathscr{F}\mathscr{A}$ and not only the usual systems are consid-

---

([3]) There are many distinct literal Gödel sentences; for example, if $\neg\exists x \mathrm{Prov}_{\mathscr{F}}(x, \sigma a)$ has Gödel number n_0, where a is a free variable and σ Gödel's substitution function, then $\neg\exists x \mathrm{Prov}_{\mathscr{F}}(x, \sigma s^{n_0} 0)$ is a literal Gödel sentence (where s is the successor symbol); but so is $\neg\exists x \mathrm{Prov}_{\mathscr{F}}(x, \sigma s^{n_1} 0)$, where n_1 is the Gödel number of $\neg\exists x \mathrm{Prov}_{\mathscr{F}}[x, (\sigma a) + 0]$.

ered, Gödel's second theorem and Löb's are *incomparable*. On the one hand, Gödel's is a special case of Löb's [11], p. 155, 3.2331, for the usual systems; on the other hand, since $\mathscr{CFA}$ is complete w.r.t. Σ_1^0 sentences, by [9] Gödel's theorem extends to $\mathscr{CFA}$, and, by above, even its sharpened form: $A_G \leftrightarrow \mathrm{Con}_{\mathscr{CFA}}$ (while Löb's does not).

Evidently, literal Henkin and Gödel sentences A_H and A_G are intended to approximate the idea of actually *expressing* their own derivability, resp. underivability; of respecting synonymity and not only formal equivalence. We cannot be sure that the concept of literal A_H and A_G is adequate for a theory of *synonymity* (between such sentences); if the subject of synonymity lends itself to theory at all! cf. III.3. But we recommend readers to keep in mind the formal facts presented in Part I when judging the uses — or abuses — of self referential propositions in the philosophical literature.

We do not know if the time is ripe to set out in full generality the properties (of $\mathscr{CFA}$) needed for the principal results in I.2; results which distinguish $\mathscr{CFA}$ from the usual systems (satisfying the conditions of A.2). At any rate we have not taken the time to do so. But we have the impression we have a pretty good idea of those properties, and so we set them out in Part II separately. From a different point of view, Part II may be seen to contain the new *general metamathematical facts* which happened to have been discovered for the sake of our results in Part I.

One group of such facts concerns the set (of formulae)

$$\boldsymbol{CF} = \{A\colon \vdash_{\mathscr{CFA}} CF(A)\}\ (4),$$

where $CF(A)$ stands for

$$(\vdash_{\mathscr{CFA}} A) \leftrightarrow (\vdash_{\mathscr{UA}} A)$$

and '*CF*' for cut free; cf. also fundamental conjecture of [22] or 'conjecture fondamentale' of [4].

The role of $\boldsymbol{CF}$ in connection with Gödel sentences A_G for $\mathscr{CFA}$ is very simple.

If for all A_G, $A_G \in \boldsymbol{CF}$, then all A_G are equivalent.

Firstly, if $A_G \in \boldsymbol{CF}$, A_G is evidently also a Gödel sentence for $\mathscr{UA}$, and hence $\vdash_{\mathscr{UA}} (A_G \leftrightarrow \mathrm{Con}_{\mathscr{UA}})$ since $\mathscr{UA}$ is a 'usual' system; equivalently since $\forall A\,[CF(A)]$ is true by [21], $\vdash_{\mathscr{CFA}} (A_G \leftrightarrow \mathrm{Con}_{\mathscr{UA}})$.

Secondly, $\mathrm{Con}_{\mathscr{UA}} \leftrightarrow \mathrm{Con}_{\mathscr{CFA}}$ can be derived (even in $\mathscr{PRA}$, that is, primitive recursive arithmetic, by II.2.2), and so

$$\vdash_{\mathscr{CFA}} (A_G \leftrightarrow \mathrm{Con}_{\mathscr{CFA}}).$$

---

(4) Actually, for many classes $\boldsymbol{A}$, there is a weak subsystem F_0 of analysis such that

$$\boldsymbol{A} \cap \boldsymbol{CF} = \boldsymbol{A} \cap \{A\colon CF(A) \text{ is derivable in } F_0\}$$

e.g. if $\boldsymbol{A}$ is the class of A_G.

Thirdly, and quite generally, for any two formulae $A, B \in \boldsymbol{CF}$, if $\vdash_{\mathscr{CFA}}(A \leftrightarrow B)$, their deductive equivalence

$$(\vdash_{\mathscr{CFA}} A) \leftrightarrow (\vdash_{\mathscr{CFA}} B)$$

can be derived in $\mathscr{CFA}$ since (even in $\mathscr{PRA}$)

$$[\vdash_{\mathscr{CFA}}(A \leftrightarrow B)] \rightarrow [\vdash_{\mathscr{A}}(A \leftrightarrow B)]$$

and

$$[\vdash_{\mathscr{A}}(A \leftrightarrow B)] \rightarrow [(\vdash_{\mathscr{A}} A) \leftrightarrow (\vdash_{\mathscr{A}} B)].$$

Thus if both A_G and $\mathrm{Con}_{\mathscr{CFA}} \in \boldsymbol{CF}$, they are formally and deductively equivalent (in $\mathscr{CFA}$).

Literal Gödel sentences (and, in particular, literal $\mathrm{Con}_{\mathscr{CFA}}$) are easily shown to $\in \boldsymbol{CF}$; cf. II.2.2. For arbitrary A_G we need a general result of Girard (proved in [4]), which supersedes — at least for the present purpose — earlier partial results which we had obtained by proof theoretic methods; for uses of these methods cf. III.2.

The other group of metamathematical facts, needed for our results on *literal* Henkin and Rosser sentences for $\mathscr{CFA}$, concerns various refinements of the result:

> *the reflection principle for Σ^0_1 sentences of $\mathscr{CFA}$ can be proved in $\mathscr{CFA}$ itself.*

(where Σ^0_1 means the class of those sentences which, demonstrably in $\mathscr{CFA}$, are formally *and* deductively equivalent to a literal Σ^0_1 formula). This complements the known result ([13]; Technical Note II, p. 136, l. 13–19) concerning the role of Σ^0_1 sentences, namely that $(\forall A \in \Sigma^0_1) CF(A)$, $\forall A\,[CF(A)]$, $(\forall A \in \Sigma^0_1)(\vdash_{\mathscr{A}} A \rightarrow A)$ are demonstrably equivalent in $\mathscr{PRA}$.

For reasons elaborated in Part III we believe that both topics of self referential propositions and of cut free analysis are rewarding in the present state of knowledge. But our paper assumes less. If a reader has doubts about the significance — or is ignorant of the details — of $\mathscr{CFA}$, he should think of our results in Part I as 'counterexamples' in a general theory of self-referential propositions of formal systems; examples that show which conditions used in the literature on self-referential propositions are necessary. If a reader thinks of such propositions as, principally, a source of formal tricks — rather than as a subject of independent interest — he should concentrate on Part II, which is relevant provided only something like $\mathscr{CFA}$ is viable; cf. III.2.

Part II is then of interest if one wants to know what can be established about $\mathscr{CFA}$ with *restricted metamathematical methods.* (The technical role of formally self-referential propositions was to be expected in this connection; after all, Gödel's incompleteness theorems are, still, the most interesting results which *obviously* involve a restriction of metamathematical

methods.) From this point of view the results of Part II give new significance to *normal form theorems* (or completeness theorems relative to $\mathfrak{U}\mathfrak{A}$, as they were called in [12], p. 329). In particular, it is not true that they are superseded by so-called *normalization theorems*, which establish that a *particular* normalization (or, equivalently, cut elimination) procedure terminates. Specifically, in Part III we find $A \in \boldsymbol{CF}$, for which it cannot be proved in analysis that the normalization procedure in question terminates when applied to an arbitrary derivation (in $\mathfrak{U}\mathfrak{A}$) of A. More generally, we question in Part III the privileged role of that procedure, which treats derivations only according to their logical *form*. In contrast, for some $A \in \boldsymbol{CF}$, we introduce different normalization procedures depending on A, that is, on the (logical or mathematical) *content* of the derivations with the end formula A, and raise a question for A_G (which $\in \boldsymbol{CF}$).

It is a pleasure to dedicate this note to Professor A. Mostowski who published the first systematic analysis [19] of Gödel's imcompleteness theory, separating some of its different aspects. Mostowski considered particularly the first incompleteness theorem (in the sense of the existence of *some* formally undecided proposition) by reference to (local) *definability* properties of truth and derivability. He thereby extended to non-formal theories (in the sense of Tarski), that is to sets of formulae closed under classical logical consequence, an earlier discovery of the purely recursion theoretic character of the *first* incompleteness theorem for formal theories, namely this: formal rules $\mathfrak{F}$ generate r.e. sets of theorems, but (there is a Π^0_1 – predicate, for example) $\{e\colon \forall x \neg T(e, e, x)\}$ for Kleene's T predicate (which) is not r.e. Thus

$$\{e\colon \forall x \neg T(e, e, x)\} \neq \{e\colon \vdash_{\mathfrak{F}} T_e\},$$

where T_e is the representation in $\mathfrak{F}$ of the proposition $\forall x \neg T(e, e, x)$ for $e = 0, 1,$

I. RESULTS ON SELF-REFERENTIAL PROPOSITIONS

As in the introduction, a formal system is called *usual* if it satisfies the conditions listed in A.2. In Section 1.1 below we introduce some refinements (of familiar metamathematical notions) needed for the study of $\mathscr{GFA}$ and of other unusual systems.

1. Definitions of some principal metamathematical notions (about formal systems) and review of results scattered in the literature on usual systems.

1.1. Canonical coding of syntactic objects (words of a finite alphabet) and of (syntactic) relations between such objects, for example the proof predicate of a formal system. The essential point about conditions C on 'canonical codings' of finite sequences, familiar from set theory or arithmetic, is that two codings satisfying C are demonstrably isomorphic; see e.g. [17]; p. 256, 2.5.3, for such conditions. The essential point about conditions on canonical codings of syntactic relations — for a given coding or 'gödelization' of the syntactic objects — is that two such codings should be demonstrably equivalent, see e.g. [11]; p. 154, 3.222, for such conditions when the formal system is presented by (a finite number of) Post rules.

It may be remarked that, for systems like $\mathscr{GFA}$, where deductive equivalence is not derivable from formal equivalence, one may sharpen the conditions on canonical codings of relations to ensure that two such codings will be also deductively equivalent. It so happens that this sharper condition will not be used in our paper.

1.2. Demonstrable completeness of $\mathscr{F}$ for closed Σ_1^0 formulae (where $\mathscr{F}$ either contains the language of arithmetic or a definitional extension): for each closed $A \in \Sigma_1^0$: $\vdash_{\mathscr{F}}(A \to \vdash_{\mathscr{F}} A)$.

We shall also need a *uniform* version, where A contains the free (numerical) variable n and the term s_A, with the variable n, defines canonically the Gödel numbers of the numerical instances of A. Thus $s_A[n/0]$, $s_A[n/s0]$, are the Gödel numbers of $A[n/0]$, $A[n/s0]$, and they are closed Σ_1^0 formulae, s being the successor symbol.

$\mathscr{F}$ is called demonstrably uniformly complete for Σ_1^0 formulae if

$$\vdash_{\mathscr{F}} \forall n [A \to \exists p \operatorname{Prov}_{\mathscr{F}}(p, s_A)].$$

Remarks. Demonstrable completeness for closed Σ_1^0 formulae is condition III in A.2. In the analysis of [8], p. 295, we have instead demonstrable uniform completeness for canonical definitions of primitive recursive P (in place of the Σ_1^0 formula A). The natural, quantifier free formulation requires a primitive recursive term π_P such that

$$\vdash_F [P \to \operatorname{Prov}_F(\pi_P, s_P)].$$

(The intended π_P defines the Gödel number of the ordinary computations of numerical instances of P.) Demonstrable completeness for the Σ_1^0 formulae $\exists n P$ follows then from (demonstrable) closure under the cut free rule $P \vdash\!\to \exists n P$, that is

$$\vdash_{\mathscr{F}} [\exists p \operatorname{Prov}_{\mathscr{F}}(p, s_P) \to \exists p \operatorname{Prov}_{\mathscr{F}}(p, e_P)],$$

where e_P defines the Gödel number of $\exists n P$. (In future, our descriptions will be less pedantic.)

1.3. Demonstrable closure of $\mathscr{F}$ under *modus ponens*. There is a *uniform* version, for variable A and B

$$\vdash_{\mathscr{F}} \{[\vdash_{\mathscr{F}}(A \to B)] \to [(\vdash_{\mathscr{F}} A) \to (\vdash_{\mathscr{F}} B)]\}$$

and a *local* version, for given A_0 and B_0, if $\vdash_{\mathscr{F}}(A_0 \to B_0)$, then

$$\vdash_{\mathscr{F}} [(\vdash_{\mathscr{F}} A_0) \to (\vdash_{\mathscr{F}} B_0)].$$

Actually for the usual systems the uniform version holds and for $\mathscr{CFA}$ the local one is refuted (by suitable A_0, B_0).

For comparison, for example with [8]: there is a different form of the local version, which assumes that B_0 is derivable from A_0 in $\mathscr{F}$, instead of assuming $\vdash_{\mathscr{F}}(A_0 \to B_0)$. For systems with quantifiers this amounts to $\vdash_{\mathscr{F}}(A_1 \to B_0)$, where A_1 is the closure of A_0.

1.3.1. Demonstrable closure under *numerical substitution*, w.r.t. the formula A, is expressed by

$$\vdash_{\mathscr{F}} [(\vdash_{\mathscr{F}} \forall n A) \to \forall n \exists p \operatorname{Prov}_{\mathscr{F}}(p, s_A)].$$

It is related to closure under *modus ponens* since for each numerical instance A_m ($m = 0, 1, \ldots$), $\vdash_{\mathscr{F}} [(\forall n A) \to A_m]$.

1.3.2. *Deductive equivalence and a counterexample.* We write

$$A \equiv_{\mathscr{F}} B$$

for $(\vdash_{\mathscr{F}} A) \leftrightarrow (\vdash_{\mathscr{F}} B)$; thus demonstrable deductive equivalence, mentioned in the introduction, is expressed by $\vdash_{\mathscr{F}} (A \equiv_{\mathscr{F}} B)$. For systems demonstrably closed under *modus ponens*

$$[\vdash_{\mathscr{F}} (A \leftrightarrow B)] \Rightarrow \vdash_{\mathscr{F}} (A \equiv_{\mathscr{F}} B).$$

Remark. The converse is, generally, false. (We prove the result for the usual system $\mathscr{U}\mathscr{A}$; but the argument also applies to $\mathscr{C}\mathscr{F}\mathscr{A}$ if we put the dummy quantifier $\forall n$ before the formula $s0 = 0$ used below since then all formulae employed satisfy

$$\vdash_{\mathscr{C}\mathscr{F}\mathscr{A}} [(\vdash_{\mathscr{U}\mathscr{A}} A) \leftrightarrow (\vdash_{\mathscr{C}\mathscr{F}\mathscr{A}} A)].$$

Take $\mathrm{Con}_{\mathscr{U}\mathscr{A}}$ for A and $s0 = 0$ or, equivalently, $\forall n(s0 = 0)$ for B. By Gödel's second theorem, $A \equiv_{\mathscr{U}\mathscr{A}} B$ can be derived in $\mathscr{U}\mathscr{A}$ (in fact, also in primitive recursive arithmetic $\mathscr{P}\mathscr{R}\mathscr{A}$), but A is true and B false, and hence $A \leftrightarrow B$ is false.

If we merely want a counter example to

$$\vdash_{\mathscr{U}\mathscr{A}} [(\vdash_{\mathscr{U}\mathscr{A}} A) \rightarrow (\vdash_{\mathscr{U}\mathscr{A}} B)] \Rightarrow [\vdash_{\mathscr{U}\mathscr{A}} (A \rightarrow B)]$$

it is enough to use Gödel's first theorem; take $A \in \Pi_1^0$, which is true but not derivable in $\mathscr{U}\mathscr{A}$, that is, $(A \rightarrow \vdash_{\mathscr{U}\mathscr{A}} A)$ is false, and $(\vdash_{\mathscr{U}\mathscr{A}} A)$ for B. By completeness of $\mathscr{U}\mathscr{A}$ for the Σ_1^0 formula $(\vdash_{\mathscr{U}\mathscr{A}} A)$, $[(\vdash_{\mathscr{U}\mathscr{A}} A) \rightarrow (\vdash_{\mathscr{U}\mathscr{A}} B)]$ is derivable, actually in $\mathscr{P}\mathscr{R}\mathscr{A}$.

1.4. Demonstrable instances of the local reflection schema

$$(\vdash_{\mathscr{F}} A) \rightarrow A \quad \text{for closed } A.$$

(See [16] for a systematic development, where also the *uniform* version is discussed.)

Without any other assumption than: $\vdash_{\mathscr{F}} A$ derivable (in $\mathscr{F}$) iff A is in fact derivable, we have, by [5]:

> if $\vdash_{\mathscr{F}} (A_G \leftrightarrow \neg \vdash_{\mathscr{F}} A_G)$, then the instance of the schema with A_G for A cannot be derivable in (a consistent) $\mathscr{F}$.

For if it were, we'd have both

> $(\vdash_{\mathscr{F}} A_G) \rightarrow A_G$ and $\neg \vdash_{\mathscr{F}} A_G \rightarrow A_G$; hence $\neg\neg A_G$ and therefore also $\neg(\neg\neg) \vdash_{\mathscr{F}} A_G$ would be derivable in $\mathscr{F}$, which would make $\mathscr{F}$ inconsistent.

(We assume here only that $\mathscr{F}$ *is* closed under (classical or intuitionistic) propositional inference, not necessarily demonstrably so. NB. In general A_G itself need not be undecided since in a consistent but ω-inconsistent $\mathscr{F}$, $\neg A_G$ may be derivable; cf. [16], p. 109.)

For $\mathscr{F}$ satisfying (1.2) and (1.3), Löb's theorem extends this result to:

For *all* A, $(\vdash_{\mathscr{F}} A)\to A$ is derivable in $\mathscr{F}$ iff A itself is so derivable.

This is an extension inasmuch as [5] asserts only that the particular formula A_G cannot be derivable.

1.5. Local and global consistency. Let $\mathrm{Con}_{\mathscr{F}}(A)$ be short for

$$(\vdash_{\mathscr{F}} A)\to\neg(\vdash_{\mathscr{F}}\neg A) \quad \text{(local at } A\text{)}$$

and let $\mathrm{Con}_{\mathscr{F}}$ stand for $\forall A\,\mathrm{Con}_{\mathscr{F}}(A)$, where the quantifier A ranges over all formulae of $\mathscr{F}$.

NB. Occasionally, attention will have to be given to the *exact* formulation of (local or global) consistency of unusual systems; cf. II. 3.1 concerning the deductive equivalence between the formulations above and $\neg[(\vdash_{\mathscr{F}} A)\wedge(\vdash_{\mathscr{F}}\neg A)]$, resp. $\neg(\exists A)\,[(\vdash_{\mathscr{F}} A)\wedge(\vdash_{\mathscr{F}}\neg A)]$.

For the usual systems $\mathscr{F}$ the distinction between local and global consistency is not interesting since all inconsistencies are formally equivalent and, hence under 1.3, deductively equivalent so that

$$\mathrm{Con}_{\mathscr{F}}(A)\to\mathrm{Con}_{\mathscr{F}}.$$

Obviously $\mathrm{Con}_{\mathscr{F}}\to\mathrm{Con}_{\mathscr{F}}(A)$ and so: $\mathrm{Con}_{\mathscr{F}}(A)\leftrightarrow\mathrm{Con}_{\mathscr{F}}$. This will be shown to be false for $\mathscr{CFA}$.

For reference in 1.6. and A.1, we restate Gödel's second theorem (on the underivability of $\mathrm{Con}_{\mathscr{F}}$). As in 1.4, suppose $\vdash_{\mathscr{F}}(A_G\leftrightarrow\neg\vdash_{\mathscr{F}} A_G)$.
Since $(\vdash_{\mathscr{F}} A_G)\in\Sigma_1^0$, by 1.2: $(\vdash_{\mathscr{F}} A_G)\to\vdash_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G)$ (is derivable in $\mathscr{F}$).
By construction of A_G and 1.3: $(\vdash_{\mathscr{F}} A_G)\to\vdash_{\mathscr{F}}(\neg\vdash_{\mathscr{F}} A_G)$, and so

$$\mathrm{Con}_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G)\to\neg\vdash_{\mathscr{F}} A_G, \quad \text{hence} \quad \mathrm{Con}_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G)\to A_G.$$

Conversely, since the underivability of any formula ensures consistency (for the usual systems) also

$$A_G\to\mathrm{Con}_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G).$$

The restatement, with $\mathrm{Con}_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G)$ in place of the traditional $\mathrm{Con}_{\mathscr{F}}$, has purely pedagogic interest *here*, since $\mathrm{Con}_{\mathscr{F}}$ and $\mathrm{Con}_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G)$ are both formally and deductively equivalent under present assumptions.

1.6. General and literal Gödel sentences 'expressing' their own underivability in $\mathscr{F}$. General Gödel sentences, A_G, have already been introduced in (1.4). They are defined by the condition

$$\vdash_{\mathscr{F}}(A_G\leftrightarrow\neg\vdash_{\mathscr{F}} A).$$

Literal Gödel sentences are required to have the form $\neg\vdash_{\mathscr{F}} A_G$, that is,

$$\neg\exists x\,\mathrm{Prov}_{\mathscr{F}}(x,\ulcorner A_G\urcorner),$$

where $\ulcorner A_G \urcorner$ is a term that canonically defines the syntactic object A_G. NB. In the case of arithmetic coding, $\ulcorner A_G \urcorner$ will *not* generally be a numeral (but another term whose value is the Gödel number of A_G).

Remark. Jeroslow [9] has made an interesting use of (a variant, described in A.1, of) *literal* Gödel sentences to extend Gödel's second theorem to systems satisfying (1.2), but not necessarily (1.3); for example to $\mathscr{CFA}$ which is obviously demonstrably complete for Σ_1^0 sentences. The essential observation used — though perhaps not mentioned very explicitly — by Jeroslow is seen by reference to the restatement of Gödel's second theorem in 1.5. If we use literal A_G, we do not require 1.3 to see that $(\vdash_{\mathscr{F}} A_G) \rightarrow \vdash_{\mathscr{F}} (\neg \vdash_{\mathscr{F}} A_G)$ can be derived in $\mathscr{F}$. Thus, completeness of $\mathscr{F}$ for Σ_0^1 sentences is enough to ensure the derivability in $\mathscr{F}$ of

$$(*) \qquad \mathrm{Con}_{\mathscr{F}}(\vdash_{\mathscr{F}} A_G) \rightarrow A_G$$

for *literal* A_G; see also A.1.

1.7. General and literal Henkin sentences, 'expressing' their own derivability. A general Henkin sentence, A_H, for $\mathscr{F}$ satisfies, by definition:

$$A_H \leftrightarrow \vdash_{\mathscr{F}} A_H \quad \text{is derivable in } \mathscr{F}.$$

Literal A_H are defined analogously to literal A_G in 1.6. As was shown by Löb (or, by a slightly different argument and using different conditions on $\mathscr{F}$, in A.2), *all A_H are derivable* (in $\mathscr{F}$). Obviously, there are many A_H which are not literal, for example $0 = 0$ (and in fact *every* theorem of $\mathscr{F}$ is an A_H).

Equally obviously, under the same conditions (satisfied by the usual systems), *no refutable formula can be a Henkin sentence because there is no (well formed) B such that $\neg \vdash_{\mathscr{F}} B$ is derivable in $\mathscr{F}$.*

1.8. *The Rosser variant of* $\mathscr{F}$, for short $\mathscr{F}_R$, is obtained from $\mathscr{F}$ by the following additional requirement on (the formal steps needed to verify) the property of being a derivation, say d: d must not only be built up according to the rules of $\mathscr{F}$, but no d' preceding d (in some ω-ordering of deductions of $\mathscr{F}$) may *contradict* d, that is, neither of the end formulae of d and d' may be the formal negation of the other.

Evidently $\vdash_{\mathscr{PRA}} \mathrm{Con}_{\mathscr{F}_R}$, where, as before, $\mathscr{PRA}$ is primitive recursive arithmetic.

Note that if $\mathscr{F}$ is consistent, $\mathscr{F}_R$ has not only the same theorems, but the same deductions as $\mathscr{F}$. So if $\mathscr{F}$ contains the set of theorems of arithmetic, so does $\mathscr{F}_R$ which then proves its own consistency. (Of course, as noted by Jeroslow [9], $\mathscr{F}_R$ cannot satisfy 1.2, that is, completeness for Σ_1^0 sentences.)

Gödel sentences for $\mathscr{F}_R$ or, equivalently by definition, *Rosser sentences* A_R for $\mathscr{F}$. Evidently, since, by 1.4, A_R is underivable in $\mathscr{F}_R$ (or, equivalently, in $\mathscr{F}$ — for consistent $\mathscr{F}$!) it is *not* true that $\vdash_{\mathscr{F}_R} (A_R \leftrightarrow \mathrm{Con}_R)$.

We have, however, the following relations between A_R and $\mathrm{Con}_{\mathscr{F}}$ which are implicit in the original papers of Gödel [5] and Rosser [20]: For the *usual* systems $\mathscr{F}$

$$\vdash_{\mathscr{PRA}}(\mathrm{Con}_{\mathscr{F}} \to A_R) \quad \text{but } \textit{not } \vdash_{\mathscr{F}}(A_R \to \mathrm{Con}_{\mathscr{F}}).$$

In other words, A_R is strictly 'weaker' than $\mathrm{Con}_{\mathscr{F}}$ (in the sense of $\mathscr{F}$).

Proof. By [20], $\vdash_{\mathscr{PRA}}(\mathrm{Con}_{\mathscr{F}} \to \neg \vdash_{\mathscr{F}} A_R)$, hence $\vdash_{\mathscr{PRA}}(\mathrm{Con}_{\mathscr{F}} \to A_R)$ and $\vdash_{\mathscr{PRA}}(\mathrm{Con}_{\mathscr{F}} \to \neg \vdash_{\mathscr{F}} \neg A_R)$.

By [5], for the usual systems $\mathscr{F}$, we do *not* have $\vdash_{\mathscr{F}}(\mathrm{Con}_{\mathscr{F}} \to \neg \vdash_{\mathscr{F}} \neg \mathrm{Con}_{\mathscr{F}})$. Thus $(\neg A_R)$ and $(\neg \mathrm{Con}_{\mathscr{F}})$ are not demonstrably deductively, and hence not formally equivalent.

Open questions (for the usual systems): Are all A_R or, at least, all literal A_R formally equivalent? Is $\mathscr{F}_R$ demonstrably closed under *modus ponens*?

By I.2.8 the two questions have negative answers for $\mathscr{F} = \mathscr{CFA}$ (except that *literal* Rosser sentences for $\mathscr{CFA}$ are equivalent).

2. Results concerning the notions of Section 1 for cut free classical analysis and related systems. 2.1 describes briefly our formal system; each subsection 2.i ($i > 1$) concerns the notion defined in 1.i.

2.1. Cut free classical analysis $\mathscr{CFA}$. Though for some (important) applications of 'cut free' systems in III.2, the exact choice of language and rules is critical, this is not so for our present purpose. To fix ideas, we shall take $\mathscr{CFA}$ to consist of the schemata (of the calculus of sequents) of impredicative second order logic with equality in [22], extended by symbols 0 (for zero), s (for successor), f_n (for the n-th primitive recursive function in some standard ω-ordering, e.g. [7]), the axioms, $\forall x \neg sx = 0$, $\forall x \forall y (sx = sy \to x = y)$, $\forall x \forall y (x = y \vee \neg x = y)$, and the 'recursion equations'

$f_n 0 = t_n$, where t_n is built up from 0, s, f_m ($m < n$), and parameters,

$\forall x [f_n(sx) = t'_n]$, where t'_n is a (suitable) term built up from 0, x, s, f_m ($m < n$), $f_n(x)$, and the same parameters as above.

Also we have a *limited cut rule*: from A and $A \to B$ derive B, for any conjunction A of the axioms above (including the recursion equations).

NB. The systems in [4] or in [5] would do equally well. We could replace the limited cut rule by suitable rules corresponding to the axioms.

Let N be short for (Dedekind's inductive definition of the natural numbers):

$$\forall X [\{X(0) \wedge \forall y [X(y) \to X(sy)]\} \to X(x)]$$

or (Zermelo's):

$$\forall X\big[\{X(x)\wedge\forall y[X(sy)\to X(y)]\}\to X(0)\big].$$

If the equality axiom $\forall X\forall u\forall v\big(u=v\to[X(u)\leftrightarrow X(v)]\big)$ is not included in (second order) logic we restrict Dedekind's definition to 'extensional' X and take for N:

$$\forall X\big[\big\{X(0)\wedge\forall y[X(y)\to X(sy)]\wedge\forall u\forall v\big(u=v\to[X(u)\leftrightarrow X(v)]\big)\big\}\to X(x)\big].$$

A formula whose variables for individuals are restricted to $\in N$, and set variables $\subset N$ is called *purely analytic*. Despite its name, not all formulae of $\mathscr{CFA}$ are purely analytic.

Restricted quantifiers $\forall x(N\to$ or $\exists x(N\wedge$ will often be replaced by $\forall n(\ \ $, resp. $\exists n($

A formula A (not the predicate defined by A!) is said to be in

$$\Sigma^0_0 \quad \text{(or, equivalently, } \Pi^0_0\text{)}, \quad \Sigma^0_1, \quad \Pi^0_1$$

if A is closed and

$$\text{quantifier free, of the form } \exists nB, \text{ resp. } \forall nB,$$

where B is quantifier free, that is, built up of the symbols 0, s, f and variables for individuals; instead of a single quantifier $\exists n$, $\forall n$ we may also have sequences (of existential, resp. universal ones).

$$\text{Thus } 0=s0\in\Sigma^0_0, \text{ but } \forall n(0=s0)\in\Pi^0_1.$$

Usual classical analysis, already mentioned in the introduction, $\mathscr{UA}$, is obtained from $\mathscr{CFA}$ by adding the full rule of cut. To simplify notation we shall use

$$\vdash^+ \text{ and } \mathrm{Con}^+ \text{ for } \vdash_{\mathscr{UA}} \text{ and } \mathrm{Con}_{\mathscr{UA}}$$

$$\vdash^- \text{ and } \mathrm{Con}^- \text{ for } \vdash_{\mathscr{CFA}} \text{ and } \mathrm{Con}_{\mathscr{CFA}}$$

unless special emphasis is required;

$$CF(A) \text{ stands for: } (\vdash^+A)\leftrightarrow(\vdash^-A).$$

We put $A\equiv B$ for deductive equivalence $(\vdash^-A)\leftrightarrow(\vdash^-B)$, simplifying the notation of 1.3.2. We shall not need deductive equivalence for $\mathscr{UA}$ (which is in any case derivable, in $\mathscr{PRA}$, from formal equivalence in case of $\mathscr{UA}$) often enough to justify a special notation.

2.2. $\mathscr{CFA}$ is, demonstrably, uniformly complete for formulae A in $\Sigma^0_0\cup\Sigma^0_1$.

In fact, more is true: $\vdash_{\mathscr{PRA}}\forall n[A\leftrightarrow\exists p\,\mathrm{Prov}^-(p,s_A)]$, which is the conjunction of uniform completeness and of the uniform reflection principle for $\Sigma^0_0\cup\Sigma^0_1$; for the proof, see II.1.a.

2.3.(a). $\mathscr{CF}\mathscr{A}$ is not (even locally) closed under *modus ponens*.

To see this let $A_0 \in \Pi_0^0$ and $B_0 \in \Pi_1^0$ and false so both $\neg A_0$ and $\neg B_0$ are true and therefore (by 2.2) derivable. Thus

$$\vdash^- [(\neg A_0) \leftrightarrow (\neg B_0)] \text{ and hence } \vdash^- (A_0 \leftrightarrow B_0).$$

But neither $\vdash^- [\vdash^- A_0) \leftrightarrow \vdash^- B_0)]$ nor, equivalently by [21], $\vdash^+ [(\vdash^- A_0) \leftrightarrow (\vdash^- B_0)]$. For, since A_0 is false and in Π_0^0, by 2.2

$$\vdash_{\mathscr{PRA}} (\neg \vdash^- A_0);$$

but not

$$\vdash^+ (\neg \vdash^- B_0)$$

since $\vdash_{\mathscr{PRA}} CF(B_0)$ and hence $\vdash^+ (\neg \vdash^- B_0) \Rightarrow \vdash^+ (\neg \vdash^+ B_0)$ which is impossible since $\mathscr{UA}$ is a usual system.

2.3.(b). $\mathscr{CF}\mathscr{A}$ is not (even) locally closed under substitution in the sense of I.1.2.1. In fact for all Π_1^0 formulae $\forall n P$

$$(\vdash^- \forall n P) \rightarrow \forall n \exists p \operatorname{Prov}^- (p, s_P)$$

s derivable in $\mathscr{CF}\mathscr{A}$ (or, equivalently, $\mathscr{UA}$) if and only if $\vdash^- \forall n P$.

Since $\forall n P \in \Pi_1^0$, by II.2.2, $\vdash_{\mathscr{PRA}} CF(\forall n P)$.

Since $P \in \Sigma_0^0$, by 2.2: $\vdash_{\mathscr{PRA}} \forall n [\exists p \operatorname{Prov}^- (p, s_P) \rightarrow P]$.

Thus $(\vdash^- \forall n P) \rightarrow \forall n \exists p \operatorname{Prov}^- (p, s_P)$ is derivable in $\mathscr{CF}\mathscr{A}$ if and only if

$$(\vdash^+ \forall n P) \rightarrow \forall n P \text{ is derivable in } \mathscr{CF}\mathscr{A}$$

and hence by Löb's theorem, $\vdash^+ \forall n P$ or, equivalently, $\vdash^- \forall n P$.

2.4. (a) The $\Sigma_0^0 \cup \Sigma_1^0$ instances of the uniform reflection principle for $\mathscr{CF}\mathscr{A}$ are derivable (in $\mathscr{PRA}$ and hence) in $\mathscr{CF}\mathscr{A}$. This is proved in II.1.b.

(b) Evidently, if $\vdash^+ CF(A)$, Löb's theorem, for A, applies to $\mathscr{CF}\mathscr{A}$. For

$$\vdash^- [(\vdash^- A) \rightarrow A] \Rightarrow \vdash^+ [(\vdash^+ A) \rightarrow A]$$

and so, by Löb's theorem for $\mathscr{UA}$ and truth of $CF(A)$,

$$\vdash^- [(\vdash^- A) \rightarrow A] \Rightarrow \vdash^- A.$$

Corollary to (a). The following proposal of a 'weak' extension (and hence also of 'the' direct extension) of Löb's theorem is *false*:

$$\vdash^- [(\vdash^- A) \rightarrow A] \Rightarrow [(\vdash^- A) \vee \vdash^- (\neg \vdash^- A)].$$

(This is evidently true for $A \in \Sigma_0^0$.) Take $A \in \Sigma_1^0$ such that

A is false, and hence not $\vdash^- A$, but $\neg \vdash^- A$ is not derivable (in $\mathscr{CF}\mathscr{A}$ nor, equivalently, $\mathscr{UA}$).

There is such an A because the complement of $\{A\colon \vdash^- A\}$, that is $\{A\colon \neg\vdash^- A\}$, $\supset \{A\colon \vdash^-(\neg\vdash^- A)\}$, both

$\{A\colon \vdash^- A\}$ and $\{A\colon \vdash^-(\neg\vdash^- A)\}$ are r.e., but $\{A\colon \vdash^- A\}$ is not recursive.

Thus, for some A, $\neg\vdash^- A$ and also $\neg\vdash^-(\neg\vdash^- A)$. However, the premise $\vdash^-[(\vdash^- A)\to A]$ holds for all $A\in\Sigma^0_1$.

2.5. Local and global consistency. (a) For $A\in\Sigma^0_0$

$$\vdash_{\mathscr{PRA}}\mathrm{Con}^-(A)$$

see [12], p. 349, and [14], p. 166.

(b) $\vdash_{\mathscr{PRA}}(\mathrm{Con}^+\leftrightarrow\mathrm{Con}^-)$. Evidently $\vdash_{\mathscr{PRA}}(\mathrm{Con}^+\to\mathrm{Con}^-)$. To show the converse it is certainly sufficient to find an A such that, for some A',

$$(*)\qquad \vdash_{\mathscr{PRA}}[\mathrm{Con}^-(A)\to\neg(\vdash^+ A')]$$

since

$$\vdash_{\mathscr{PRA}}[(\neg\vdash^+ A')\to\mathrm{Con}^+].$$

Any true Σ^0_1 formula A satisfies $(*)$ above, if we take $\neg A$ for A'. For, by completeness of $\mathscr{CFA}$ for Σ^0_1 formulae, $\vdash^- A$ and hence

$$\vdash_{\mathscr{PRA}}[(\mathrm{Con}^- A)\to\neg(\vdash^-\neg A)].$$

Also, as in 2.3, $\vdash_{\mathscr{PRA}} CF(\neg A)$ since $\neg A$ is the negation of a Σ^0_1 formula and hence $(*)$.

Since both Con^+ and Con^- are in Π^0_1 we also have $\vdash_{\mathscr{PRA}}(\mathrm{Con}^+\equiv\mathrm{Con}^-)$.

Remarks. (i) Since $\mathscr{PRA}$ is quantifier free and, for example, Con or $\neg\vdash A$ are not, some of the formulae above are strictly speaking not well formed. We mean by $\mathscr{PRA}$ its *conservative extension* obtained by adding first order logic. By the original, so-called 'finitary' version of Herbrand's theorem we have a uniform (primitive recursive) method of converting any proof in 'our' $\mathscr{PRA}$ of

$$\forall nP\to\forall mQ \qquad (P, Q \text{ quantifier free})$$

into

$$\vdash_{\mathscr{PRA}}(P[n/\pi]\to Q) \qquad \text{in 'usual' } \mathscr{PRA}$$

for a suitable function term π containing the variable m and the parameters of $(\forall nP)\to(\forall mQ)$.

(ii) The simple fact observed in (a) above that $\mathrm{Con}^-(A)$ is derivable in $\mathscr{PRA}$ for Σ^0_1 formulae A (or even the special case where A is $0 = s0$) implies immediately the well-known consequence of the fundamental conjecture for $\mathscr{UA}$:

$$\vdash_{\mathscr{PRA}}(\forall A\,[CF(A)]\to\mathrm{Con}^+) \qquad \text{and, in fact,} \qquad \vdash_{\mathscr{PRA}}[CF(0 = s0)\to\mathrm{Con}^+]$$

since $\vdash^-(\neg 0 = s0)$, and $\vdash_{\mathscr{PRA}}[\neg\vdash^-(0 = s0)]$ and hence $\vdash_{\mathscr{PRA}}[CF(0 = s0)\rightarrow$ $\rightarrow\neg\vdash^+(0 = s0)]$.

(iii) For the reader who wants some exercises concerning the behavior of $\mathrm{Con}^-(A)$ for different A, we consider the subclasses of Σ_1^0 and Π_1^0; where $\mathscr{PRA} \subset \mathscr{F} \subset \mathscr{UA}$:

$$\{A\colon \vdash_{\mathscr{F}} \mathrm{Con}^-(A)\} \cap \Sigma_1^0 \text{ and } \{A\colon \vdash_{\mathscr{F}} \mathrm{Con}^-(A)\} \cap \Pi_1^0.$$

NB. These classes are *not* invariant for all $\mathscr{F}$ considered: we state the precise results for the extremes $\mathscr{PRA}$ and $\mathscr{UA}$.

(a) Suppose $A \in \Sigma_1^0$. Then

$$\vdash^+ \mathrm{Con}^-(A) \Leftrightarrow \vdash^+ \neg A, \quad \vdash_{\mathscr{PRA}} \mathrm{Con}^-(A) \Leftrightarrow \vdash_{\mathscr{PRA}} (\neg A),$$
$$A \Rightarrow \neg\vdash^+ \mathrm{Con}^-(A).$$

Since $A \in \Sigma_1^0$, $\vdash_{\mathscr{PRA}}(A \leftrightarrow \vdash^- A)$ and hence $\vdash_{\mathscr{PRA}}\{\mathrm{Con}^-(A) \leftrightarrow [(\vdash^- \neg A)$ $\rightarrow \neg A]\}$. By Löb's theorem applied to $\mathscr{UA}$, since $\vdash_{\mathscr{PRA}} CF(\neg A)$, $\vdash^+ \mathrm{Con}^-(A)$ $\Leftrightarrow \vdash^-(\neg A)$. Similarly (since Löb's theorem applies also to $\mathscr{PRA}$)

$$\vdash_{\mathscr{PRA}}[(\vdash^- \neg A) \rightarrow \neg A] \Rightarrow \vdash_{\mathscr{PRA}}[(\vdash_{\mathscr{PRA}} \neg A) \rightarrow \neg A],$$

hence

$$\vdash_{\mathscr{PRA}} \mathrm{Con}^-(A) \Rightarrow \vdash_{\mathscr{PRA}} (\neg A);$$

and $(\vdash_{\mathscr{PRA}} \neg A) \Rightarrow \vdash_{\mathscr{PRA}} \mathrm{Con}^-(A)$ is immediate.

$\vdash_{\mathscr{PRA}}\{[A \wedge \mathrm{Con}^-(A)] \rightarrow \neg\vdash^- \neg A\}$ by completeness for Σ_1^0 formulae. Since $\vdash_{\mathscr{PRA}} CF(\neg A)$, we have $\vdash_{\mathscr{PRA}}([A \wedge \mathrm{Con}^-(A)] \rightarrow \neg\vdash^+ \neg A)$ and so $\vdash_{\mathscr{PRA}}([A \wedge \mathrm{Con}^-(A)] \rightarrow \mathrm{Con}^+)$; since $A \Rightarrow \vdash_{\mathscr{PRA}} A$ and $\neg\vdash^+ \mathrm{Con}^+$, we have $A \Rightarrow \neg\vdash^+ \mathrm{Con}^-(A)$.

(b) Suppose $A \in \Pi_1^0$. Then $\vdash^+ \mathrm{Con}^-(A) \Leftrightarrow \vdash^+ A$, $\vdash_{\mathscr{PRA}} \mathrm{Con}^-(A) \Leftrightarrow \vdash_{\mathscr{PRA}} A$ and $\neg A \Rightarrow \neg\vdash^+ \mathrm{Con}^-(A)$.

The argument is dual to that in (a), using Löb's theorem for $\mathscr{UA}$ and $\mathscr{PRA}$.

2.6. Gödel sentences A_G satisfying: $\vdash^-(A_G \leftrightarrow \neg\vdash^- A_G)$. For all Gödel sentences:

$$\vdash^-[(A_G \leftrightarrow \mathrm{Con}^-) \wedge (A_G \equiv \mathrm{Con}^-)]$$

and hence

$$\vdash^-[(A_G \leftrightarrow \mathrm{Con}^+) \wedge (A_G \equiv \mathrm{Con}^+)].$$

For all *literal Gödel* sentences $\vdash_{\mathscr{PRA}}[(A_G \leftrightarrow \mathrm{Con}^-) \wedge (A_G \equiv \mathrm{Con}^-)]$.

LEMMA. *It is sufficient to prove* $\vdash CF(A_G)$, *where* $\vdash$ *means* $\vdash^-$ *in the case of general Gödel sentences, and* $\vdash_{\mathscr{PRA}}$ *for literal ones. For*

$$[CF(A_G) \wedge (A_G \leftrightarrow \neg\vdash^- A_G)] \rightarrow (A_G \leftrightarrow \neg\vdash^+ A_G)$$

that is, A_G is also a Gödel sentence for $\mathscr{UA}$ (5).

(5) As observed by Girard, every Gödel sentence of $\mathscr{UA}$ is also one of $\mathscr{CFA}$.

By 1.5, we then have $\vdash_{\mathscr{PRA}}(A_G \leftrightarrow \mathrm{Con}^+)$. Now apply 2.5(b).)

To get $A_G \equiv \mathrm{Con}^+$ (and hence, by 2.5, also $A_G \equiv \mathrm{Con}^-$ observe that $CF(\mathrm{Con}^+)$ since $\mathrm{Con}^+ \in \Pi_1^0$, or use the assumption $CF(A_G)$ since Con^+ is certainly a Gödel sentence.

Now, to prove the principal assertions above for arbitrary A_G, we use Girard's general result, II.2.1; this gives $\vdash^- CF(A_G)$, since A_G is true, hence does not imply a false Σ_1^0 sentence, and A_G is demonstrably equivalent to the purely analytic formula $\neg\vdash^- A_G$.

For *literal* A_G, $CF(A_G)$ is proved in I.2.2 (much more simply), since such A_G are negations of Σ_1^0 sentences.

2.7. Henkin sentences A_H satisfying: $\vdash^-(A_H \leftrightarrow \vdash^- A_H)$. Evidently in constrast to the case of $\mathscr{U}\mathscr{A}$ some A_H are derivable in $\mathscr{CFA}$, some refutable; e.g. $\vdash^- 0 = 0$ but $\vdash^- \neg 0 = s0$ though both $0 = 0$ and $0 = s0$ are Henkin sentences for $\mathscr{CFA}$.

All literal Henkin sentences for $\mathscr{CFA}$ are refutable in $\mathscr{CFA}$.

Proof. It is enough to show $\vdash^-(\neg\vdash^- A_H)$ since $\vdash^-[\neg\vdash^- A_H) \leftrightarrow (\neg A_H)]$. Suppose $\vdash^- A_H$ or, more fully, suppose $\vdash^- \exists x[N(x) \wedge \mathrm{Prov}^-(x, \ulcorner A_H \urcorner)]$. Then there is a least Gödel number, n_0, of derivations of A_H, namely, the smallest numerical value such that $\mathrm{Prov}^-(n_0, \ulcorner A_H \urcorner)$. So by II.1b, with $\mathrm{Prov}^-(a, \ulcorner A_H \urcorner)$ in place of $P(a)$, there are at least n_0 logical inferences in any derivation $\vdash A_H$. This contradicts the fact that any such derivation has Gödel number $> n_0$.

Note that the argument above can be formalized in $\mathscr{PRA}$. In other words, for *literal* A_H we have

$$\vdash_{\mathscr{PRA}}(\neg A_H).$$

2.8. *Rosser variant $\mathscr{CFA}_R$ of $\mathscr{CFA}$.* As in 1.8, we call Gödel sentences of $\mathscr{CFA}_R$ also: Rosser sentences (of $\mathscr{CFA}$).

2.8.1. *For each literal Rosser sentence A_R of $\mathscr{CFA}$*

$$\vdash_{\mathscr{PRA}}(A_R \leftrightarrow \mathrm{Con}^+).$$

Proof. By definition A_R has the form $\neg\exists n \mathrm{Prov}_R^-(n, \ulcorner A_R \urcorner)$, where Prov_R^- is the (quantifier free) canonical respresentation of the proof predicate for $\mathscr{CFA}_R$. Since A_R is itself a negation the formula $\exists n \mathrm{Prov}_R^-(n, \ulcorner A_R \urcorner)$ itself contradicts A_R in the sense of 1.8. Write $\neg A_R'$ for A_R.

Now A_R' is, demonstrably in $\mathscr{PRA}$, equivalent to

$$\exists n[\mathrm{Prov}^-(n, \ulcorner A_R \urcorner) \wedge (\forall m < n) \neg \mathrm{Prov}^-(m, \ulcorner A_R' \urcorner)]$$

and so A_R is (demonstrably) equivalent to

$$\forall n[\mathrm{Prov}^-(n, \ulcorner A_R \urcorner) \rightarrow (\exists m < n)\mathrm{Prov}^-(m, \ulcorner A_R' \urcorner)].$$

But since $A'_R \in \Sigma^0_1$, we have, demonstrably,

$$A_R \leftrightarrow \forall n\,[\mathrm{Prov}^-(n, \ulcorner A_R \urcorner) \rightarrow A'_R]$$

and so

$$A_R \rightarrow \forall n \neg \mathrm{Prov}^-(n, \ulcorner A_R \urcorner) \quad \text{since } A_R \leftrightarrow \neg A'_R.$$

But since A_R is of the form $\forall n P$, by II.2.2, we have $\vdash_{\mathscr{PRA}} CF(A_R)$ and so $\vdash_{\mathscr{PRA}} (A_R \rightarrow \mathrm{Con}^+)$. The proof of the converse is classical; see 1.8.

We thus have a striking contrast to the case of $\mathscr{U}\mathscr{A}$, where we do not have equivalence between any A_R and Con^+, and where we do not know if (even) all literal A_R are equivalent.

Remark. To locate the specific step of the argument in 1.8 (showing that, for usual $\mathscr{F}$, A_R is weaker than $\mathrm{Con}_{\mathscr{F}}$) which fails for $\mathscr{CFA}$, recall that:

$$\text{not} \vdash_{\mathscr{F}} [\mathrm{Con}_{\mathscr{F}} \rightarrow \neg \vdash_{\mathscr{F}} (\neg \mathrm{Con}_{\mathscr{F}})].$$

This is used in the argument, but is false when $\mathscr{F}$ is $\mathscr{CFA}$. For $\neg \mathrm{Con}^-$ is, formally and deductively, equivalent to a Σ^0_1 formula, say A. (This is shown in II.3.1.) By 2.2

$$\vdash^- (A \leftrightarrow \vdash^- A),$$

and so by the equivalences mentioned,

$$\vdash^- [(\neg \mathrm{Con}^-) \leftrightarrow \vdash^- (\neg \mathrm{Con}^-)] \quad \text{hence} \quad \vdash^- [\mathrm{Con}^- \leftrightarrow \neg \vdash^- (\neg^- \mathrm{Con}^-)].$$

2.8.2. *There is a Rosser sentence for $\mathscr{CFA}$ which is strictly weaker than* Con^+ (and hence not all Rosser sentences for $\mathscr{CFA}$, that is, not all Gödel sentences for $\mathscr{CFA}_R$, are demonstrably equivalent in $\mathscr{CFA}_R$).

Corollary (cf. Introduction). *$\mathscr{CFA}_R$ is not demonstrably closed under modus ponens.*

Proof (of 2.8.2). Let $\tilde{A}_R$ be $\neg A^*_R$, where A^*_R is, literally,

$$\forall n \exists d\,[\mathrm{Prov}^-(d, \ulcorner \tilde{A}_R \urcorner) \wedge (\forall d' < d) \neg \mathrm{Prov}^-(d', \ulcorner A^*_R \urcorner)]$$

(where the vacuous numerical quantifier $\forall n$ is introduced to ensure $\vdash_{\mathscr{PRA}} CF(A^*_R)$ by II.2.2).

(i) $\tilde{A}_R$ is a Rosser sentence for $\mathscr{CFA}$, that is

$$(\neg A^*_R) \leftrightarrow \neg \exists d\,[\mathrm{Prov}^-(d, \ulcorner \tilde{A}_R \urcorner) \wedge (\forall d' < d) \neg \mathrm{Prov}^-(d', \ulcorner A^*_R \urcorner)]$$

is derivable in $\mathscr{CFA}$. This is seen by the truth of the fundamental conjecture (or can be verified directly; cf. (ii) below) since, in predicate logic with cut, $(\neg G) \leftrightarrow (\neg F)$ is a consequence of

$$G \leftrightarrow \forall x\,[N(x) \rightarrow F] \quad \text{and} \quad \exists x N(x)$$

provided x does not occur free in F. Since A_R^* is of the form $\forall x[N(x)\rightarrow \rightarrow F_R]$ and $\vdash_{\mathscr{CFA}}\exists x N(x)$, we get the desired result by taking A_R^* for G. (F_R is obtained from A_R^* by suppressing '$\forall n$'.)

(ii) $\vdash_{\mathscr{PRA}}(\mathrm{Con}^-\rightarrow\neg\vdash^- A_R^*)$.

As in Rosser's argument quoted in I.1.8, we have (in $\mathscr{PRA}$)

$$[\mathrm{Con}^- \wedge \mathrm{Prov}^-(d', \ulcorner A_R^*\urcorner)]\rightarrow\vdash_{\mathscr{CFA}}(\neg F_R).$$

To establish (ii) it is enough to show

$$[\vdash_{\mathscr{CFA}}(\neg F_R)]\rightarrow[\vdash_{\mathscr{CFA}}(\neg\forall n F_R)]$$

since $\neg\forall n F_R$ *is* $\neg A_R^*$. Now any derivation d (in $\mathscr{CFA}$) of $\neg F_R$ has the form

$$\begin{array}{c} d_1 \\ F_R\vdash\emptyset \\ \neg F_R \end{array}$$

cf. II.2.2. Also we have a derivation of the form

$$\begin{array}{c} d_2 \\ \neg N(0)\vdash\emptyset \end{array}$$

and hence

$$\begin{array}{cc} d_1 & d_2 \\ F_R\vdash\emptyset & \neg N(0)\vdash\emptyset \\ \multicolumn{2}{c}{N(0)\rightarrow F_R\vdash\emptyset\,.} \end{array}$$

Weakening the *LHS* by adding $\forall x[N(x)\rightarrow F_R]$ and contracting, we have a derivation of $\forall x[N(x)\rightarrow F_R]\vdash\emptyset$ and hence of $\neg A_R^*$.

(iii) $\vdash_{\mathscr{PRA}}(\mathrm{Con}^+\rightarrow\neg\vdash^+ A_R^*)$; by (ii) and $\vdash_{\mathscr{PRA}} CF(A_R^*)$.

(iv) $\tilde{A}_R\rightarrow\mathrm{Con}^-$ can*not* be derived in $\mathscr{CFA}$. If it could, so could $(\neg\mathrm{Con}^+)\rightarrow(\neg A_R)$ and also, since $\mathscr{UA}$ is demonstrably closed under cut, $[\vdash^+(\neg\mathrm{Con}^+)]\rightarrow(\vdash^+ A_R^*)$. Thus, by (iii)

$$\mathrm{Con}^+\rightarrow\neg\vdash^+(\neg\,\mathrm{Con}^+)$$

would be derivable in $\mathscr{CFA}$ and hence in $\mathscr{UA}$, contradicting Gödel's second theorem applied to (the 'usual' system consisting of) $\mathscr{UA}$ with the additional axiom Con^+; cf. I.1.8.

(v) $\vdash_{\mathscr{PRA}}(\mathrm{Con}^-\rightarrow\tilde{A}_R)$.

One shows first $\vdash_{\mathscr{PRA}}(\mathrm{Con}^-\rightarrow\neg\vdash^-\tilde{A}_R)$ as in (ii) by use of $CF(\tilde{A}_R)$ and observes $\vdash_{\mathscr{PRA}}[(\neg\vdash^-\tilde{A}_R)\rightarrow\tilde{A}_R]$.

(iv) and (v) established that $\tilde{A}_R$ is strictly weaker, in the sense of $\mathscr{CFA}$, than any *literal* Rosser sentences of $\mathscr{CFA}$ (which are equivalent to Con^-).

II. FORMALIZED METAMATHEMATICS OF $\mathscr{CFA}$

As mentioned in the introduction, we collect here some of the general results (of interest independently of our topic of self referential propositions) which were used in Part I. A word on the metamathematical methods used is in order here. At the beginning of our subject it was, perhaps, not possible to foresee what restrictions (if any) on the metamathematical methods were relevant — and doctrinaire and/or vague requirements, such as restriction to $\mathscr{PRA}$ or to 'finitary' methods, became traditional. The time has come, we are convinced, to analyze *which* restrictions are relevant (to specific results). This is of course a quite separate problem from using the 'weakest possible' methods.

Specifically, for most applications in Part I, all that matters is that the metamathematical results below are proved in $\mathscr{CFA}$ itself. An exception is the result in 1.2.7 on *literal* Henkin sentences, in the following sense: it is not (it seems) sufficient to have a proof in $\mathscr{CFA}$ of the reflection principle for Σ_1^0 sentences as formulated in I.1.4, but we need *explicit bounds*

$$(*) \qquad \forall p\,[\mathrm{Prov}_{\mathscr{CFA}}(p, \ulcorner \exists n P \urcorner) \rightarrow (\exists n < \lambda p) P]$$

for a suitable function λ of (some structural features of the derivation with Gödel number) p. The *natural* proof of $(*)$ is in $\mathscr{PRA}$, while the natural proofs of most other results seem to use the model theoretic methods of Girard. For some possible by-products of proofs in $\mathscr{PRA}$ see III.2.

1. Completeness and reflection principles for closed $\Sigma_0^0 \cap \Sigma_1^0$ formulae. For $A \in \Sigma_0^0 \cup \Sigma_1^0$

$$\vdash_{\mathscr{CFA}} (A \leftrightarrow \vdash_{\mathscr{CFA}} A).$$

(a) Demonstrable completeness of $\mathscr{CFA}$ for closed $\Sigma_0^0 \cup \Sigma_1^0$ formulae, that is

$$\vdash_{\mathscr{CFA}} (A \rightarrow \vdash_{\mathscr{CFA}} A).$$

We know, and this is spelled out in detail, e.g., in [8], that in first order arithmetic Z

$$\vdash_Z \forall n\,[P \rightarrow \exists p\, \mathrm{Prov}_Z(p, s_P)],$$

where, as in I.1.2, P is (the canonical representation of) a primitive recursive predicate with argument n, and s_P the (canonical definition of

the) Gödel number of its n-th numerical instance $P[n/0]$, $P[n/s0]$, Also — by a cut free inference — $\exists nP$ follows from any such instance; hence

$$\vdash_Z \forall p \forall n[\mathrm{Prov}_Z(p, s_P) \to \exists q \mathrm{Prov}_Z(q, \ulcorner \exists nP \urcorner)].$$

Now Z is, demonstrably, contained — via the mapping of [22] — in a *fragment* of usual analysis $\mathscr{U}\mathscr{A}$: the Π^1_1-fragment of $\mathscr{U}\mathscr{A}$ [23] is certainly sufficient. But for such a fragment the normal form theorem can be *proved* in $\mathscr{U}\mathscr{A}$ and hence, by the truth of the normal form theorem, in $\mathscr{CFA}$.

Since the argument is uniform for all closed $A \in \Sigma^0_1$, we may take, in particular, a Σ^0_1 formula A with a numerical parameter, say q, and let s_A define the q-th numerical instance of A. Then

$$\vdash_{\mathscr{CFA}} \forall q[A \to \exists p \mathrm{Prov}_{\mathscr{CFA}}(p, s_A)].$$

Remark. It is, of course, a legitimate question to ask, for primitive recursive P: How long is the shortest derivation of $P[n/0], \ldots, P[n/s^k 0], \ldots$? (that is, of the k-th numerical instance of P, as a function of k), whenever $P[n/0], \ldots$ is true. Does the fact that we are not allowed to use *modus ponens* lengthen the derivation of numerical formulae inordinately? (From I.2.3(b) we see that proofs of numerical instances may, loosely speaking, have to be much longer than the shortest proof of the general statement.) The natural answer to this quantitative question uses $\mathscr{PRA}$.

(b) Reflection principles for closed $\Sigma^0_0 \cup \Sigma^1_0$ sentences. We need the sharper form

$$\vdash_{\mathscr{CFA}} [\mathrm{Prov}_{\mathscr{CFA}}(p, \ulcorner \exists nP \urcorner) \to (\exists n < \lambda p) P],$$

where λp is the *number* of logical inferences in the derivation (coded by) p.

Remark. For $\exists xP$ in place of $\exists nP$, that is, of $\exists x(N \wedge P)$, we still have a proof in $\mathscr{PRA}$ of

$$(*) \qquad \mathrm{Prov}_{\mathscr{CFA}}(p, \ulcorner \exists xP \urcorner) \to \exists nP;$$

and, in fact, for every first order formula A

$$\vdash_{\mathscr{CFA}} [\mathrm{Prov}_{\mathscr{CFA}}(p, \ulcorner A \urcorner) \to A^N],$$

where A^N is the 'relativization' of A to N; this is the reflection principle for predicate logic which can be proved in first order arithmetic. But $(*)$ would not be true with $(\exists n < \lambda p)P$ in place of $\exists nP$.

It will be convenient to reduce result (b) to a familiar fact about cut free first order logic, which removes the apparent mystery: though $\mathscr{CFA}$ contains symbols for all primitive recursive functions and therefore short expressions with large numerical values, nevertheless the length of a derivation of $\exists nP$ in $\mathscr{CFA}$ is bounded (below) by the *numerical* value of the least realization of P.

LEMMA 1. *Let U and U' be purely universal formulae such that the instances of U are disjoint from those of U' and let P_i and Q_i be quantifier free. If d is a cut free derivation of*

$$U \vdash (U' \to Q_1) \wedge P_1, \ldots, (U' \to Q_k) \wedge P_k,$$

then there is a cut free derivation d' of

$$U \vdash \bigwedge (U' \to Q_1) \wedge P_1, \ldots, (\bigwedge U' \to Q_k) \wedge P_k,$$

where $\bigwedge U'$ is a conjunction of quantifier-free instances of U' and the number of conjuncts is less than the number of logical inferences in d.

To define $\bigwedge U'$ we note first that the only inferences in d for which either U' or $U' \to Q_i$ is a *principal* formula are

(i) the introduction of $(U' \to Q_i) \wedge P_i$ on the *RHS*,

(ii) an inference of $\Gamma \vdash U' \to Q_i$ from $\Gamma,\ U' \vdash Q_i$,

(iii) an inference of $\Gamma',\ U' \vdash \Delta$ from $\Gamma',\ U'_j,\ U' \vdash \Delta$, where U'_j is an instance of U' (introduction of a universal quantifier on the left).

We take for $\bigwedge U'$ the conjunction of *all* U'_j in the inferences in d of type (iii), and replace U' everywhere in d by $\bigwedge U'$ The resulting tree of formulae is still a cut free derivation, for the three types of inference (i)–(iii) remain valid, and no side condition for the introduction of other quantifiers is violated, the only quantifiers occurring in d being negative and universal (in U). Clearly the bound for the length of $\bigwedge U'$ holds since each U'_j arises from a separate logical inference (applied to U' on the left) and in addition there must be some inferences of type (ii), even if the Q_i are absent.

Remark. The lemma is a slight modification of the familiar uniformity theorem, where we drop U and P_i and take Q_i to be false: from a cut free derivation of $U' \vdash \emptyset$ we obtain one of $\bigwedge U' \vdash \emptyset$ as above. To avoid errors the reader should recall the need for the restriction to purely universal U'; for example the analogue would be false if we took

$$\forall x [\neg P(x) \wedge \forall y P(y)]$$

for U' and substitution instances $\neg P(t_j) \wedge \forall y P(y)$ for U'_j. (The argument above would not apply since the side conditions on quantifiers would be violated.)

LEMMA 2. *Let Γ_0 be a finite subset of the non-logical axioms of $\mathscr{CFA}$ and of the equality axioms, and let the (inductive) definition N in* I.2.1 *(of the property of being a natural number) be put in the form $\forall X N_1$. Suppose d is a cut free (first order) derivation of*

$$\Gamma_0 \vdash N_1[x/t_1] \wedge P[x/t_1], \ldots, N_1[x/t_k] \wedge P[x/t_k],$$

where the t_i are closed terms built up from the symbols in Γ_0, and let P be also a quantifier free expression built up from these symbols. Then the number of logical inferences in d exceeds the least numerical value $|t_i|$ $(1 \leqslant i \leqslant k)$ which satisfies P for the intended interpretation of the symbols.

To apply Lemma 1, we note that Γ_0 is universal and that $N_1[x/t_i]$ has the form

$$\{\forall x \forall y (x = y \to [X(x) \to X(y)]) \wedge X(0) \wedge \forall x [X(x) \to X(sx)]\} \to X(t_i).$$

Taking Γ_0 for U and the expression $\{\ \}$ for U' we note that their instances are disjoint since X does not occur in U; we take $X(t_i)$ for Q_i and $P[x/t_i]$ for P_i.

Consider then the set of values $|t|$ of the terms occurring in $\bigwedge U'$ as defined in Lemma 1 and let n_0 be the smallest number which does not occur in this set. It is clearly $<$ the total number of inferences in d. We interpret the symbols in Γ_0 in the intended way, take X to be true for $n \leqslant n_0$ and false for $n > n_0$. Then $\bigwedge U'$ is satisfied too because, for all instances $X(t) \to X(st)$ which occur in $\bigwedge U'$, $|t| \neq n_0$ by hypothesis. But then Q_i is false for all i for which $|t_i| \geqslant n_0$ and so some P_j with $|t_j| < n_0$ must be true.

Remark. We assumed that all terms are closed and built up from the symbols of Γ_0. The lemma extends obviously by use of some conventional valuation $|t|$; for example, give the value 0 to free variables and 0 to function variables not occurring in Γ_0. Obviously the lemma does not extend in general to, say, universal P_i; cf. the remark at the end of Lemma 2.

Proof of theorem. We now consider, again, a quantifier free P, but a (cut free) derivation d_2 of the sequent

$$\Gamma_0 \vdash \exists n P \quad \text{that is} \quad \Gamma_0 \vdash \exists x (P \wedge \forall X N_1)$$

with the single second order quantifier $\forall X$ on the *RHS*.

We shall literally reduce d_2 to a (shorter) first order derivation d of

$$\Gamma_0 \vdash N_1[x/t_1] \wedge P[x/t_1], \ldots, N_1[x/t_k] \wedge P[x/t_k]$$

(as in Lemma 2) by essential use of the subformula property for $\forall X N_1$: since its sole second order quantifier is universal and occurs positively, its immediate subformulae are N_1 itself up to renaming of its (sole) second order variable X. Let d_2' be obtained from d_2 by replacing $\forall X N_1$ by N_1 itself and also replacing *all* free second order variables by X. Then d_2' is also a cut free derivation, of $\Gamma_0 \vdash \exists x (P \wedge N_1)$, except that those lines of d_2 are repeated where $\forall X N_1$ is inferred from some notational variant of N_1. No side condition on the introduction of quantifiers can be violated because $\forall X$ is the *only* second order quantifier in d_2.

The formulae $\exists x(P \wedge N_1)$ can only occur on the *RHS* of any sequent in d_2' because it is not a subformula of Γ_0. Let Δ be the set of all subformulae of the form $P[x/t] \wedge N_1[x/t]$ or $N_1[x/t] \wedge P[x/t]$ which occur on the *RHS* of any sequent of d_2'. Their number is clearly bounded by the number of inferences in d_2' and *a fortiori* in d_2. We form d by replacing each occurrence of $\exists x(P \wedge N_1)$ in d_2' by Δ. Since the only inference rule for which $\exists x(P \wedge N_1)$ is a principal formula is:

from $\Gamma' \vdash P[x/t] \wedge N_1[x/t]$, $\exists x(P \wedge N_1)$, Δ' infer $\Gamma' \vdash \exists x(P \wedge N_1)$, Δ' and $P[x/t] \wedge N_1[x/t]$ is a member of Δ, we also have the rule:

from $\Gamma' \vdash P[x/t] \wedge N_1[x/t]$, Δ, Δ' infer $\Gamma' \vdash \Delta$, Δ' and so d is a derivation.

The proof of the theorem is complete because Lemma 2 applies to d.

Remark. Inspection of the argument shows that we have (even)

$$\vdash_{\mathscr{PRA}} [\mathrm{Prov}_{\mathscr{CFA}}(p, \ulcorner \exists n P \urcorner) \to (\exists n < \lambda p) P]$$

though we need only the 'weaker' assertion with $\vdash_{\mathscr{CFA}}$ in place of $\vdash_{\mathscr{PRA}}$. But, as so often, the fully explicit (quantifier-free) 'positive' metamathematical statement (asserting derivability, not underivability!) also has an elementary proof. — Below we use $\vdash^+$ and $\vdash^-$ in the sense explained in I.2.1.

2. Demonstrable instances of the normal form theorem. *CF*: demonstrable instances of the normal form theorem.

Note first that, for $A \in \Sigma_0^0 \cup \Sigma_1^0$: $A \in \boldsymbol{CF} \Leftrightarrow A$ is true. For, by the last section, $\vdash_{\mathscr{CFA}} [(\vdash^- A) \to A]$ and so, since $CF(A)$ means $(\vdash^+ A) \leftrightarrow [\vdash^- A)$, we have, in $\mathscr{CFA}$,

$$CF(A) \to [(\vdash^+ A) \to A].$$

But, by Löb's theorem, if $\vdash^- [(\vdash^+ A) \to A)$ and *a fortiori*

$$\vdash^+ [(\vdash^+ A) \to A], \quad \text{then } \vdash^+ A.$$

Conversely, by demonstrable completeness for Σ_1^0 sentences, if A is true, we have $\vdash^- A$ and so *a fortiori* $CF(A)$ (of course, by the *truth* of CF, for each A actually derivable in $\mathscr{UA}$, $CF(A)$ can be verified by computation.)

Put differently, $[(\forall A \in \Sigma_1^0) CF(A)] \to (\forall A \in \Sigma_1^0) [(\vdash^+ A) \to A]$ is derivable in $\mathscr{PRA}$, and so by Technical Note II of [13]

$$[(\forall A \in \Sigma_1^0) CF(A)] \to \forall A [CF(A)].$$

2.1. The singular role of Σ_1^0 sentences just noted, is beautifully enhanced by

Girard's theorem (for all A formally equivalent to a purely analytic formula): *Either, for some false* Σ^0_1 *sentence* E

$$\vdash_{\mathscr{CFA}} (A \to E)$$

or, for all A':

$$[\vdash_{\mathscr{CFA}} (A \leftrightarrow A')] \Rightarrow \vdash_{\mathscr{CFA}} CF(A') \ (^6).$$

In other words, A has to be 'very' false if $CF(A)$ cannot be proved in $\mathscr{CFA}$. Indeed, even then $CF(A)$ may be so proved, only for some A', formally equivalent to A, not $\vdash_{\mathscr{CFA}} CF(A')$. — This remark can be made more precise by reference to 2.2.

2.2. Girard's theorem gives information about derivability of $CF(A)$ in terms of *deductive properties* of A. For some applications, for example to *literal* Gödel sentences (or of course to Σ^0_1 sentences) it is useful to relate the derivability of $CF(A)$ to *syntactic properties* of A. We give an example (chosen because of its application to literal A_G). A simpler example which illustrates the same idea (adapted to the formally intuitionistic versions of $\mathscr{UA}$ and $\mathscr{CFA}$) is worked out in III.2.

Example. If A has the form $\forall n A'$ or $\neg \exists n B'$, that is

$$\forall x (N \to A'), \quad \text{resp.} \quad \neg \exists x (N \wedge B')$$

in the notation of I.2.1, where A' and B' are arbitrary and n is a — possibly dummy — numerical variable, then

$$\vdash_{\mathscr{PRA}} CF(A).$$

Proof. The details depend of course on the exact formulation of $\mathscr{CFA}$. We need only this. The rules (i)–(iv) are included in $\mathscr{CFA}$ (or, at least, $\mathscr{CFA}$ can be proved in $\mathscr{PRA}$ to be closed under these rules).

(i) Derive $\Gamma \vdash P \to Q$ from $\Gamma, P \vdash Q$ and $\Gamma \vdash \neg P$ from $\Gamma, P \vdash \emptyset$.

(ii) Provided x does not occur (free) in Γ derive $\Gamma \vdash \forall x F$ from $\Gamma \vdash F$ and $\Gamma, \exists x F \vdash \emptyset$ from $\Gamma, F \vdash \emptyset$.

(iii) Derive $\Gamma, P \to Q \vdash \Delta$ from $\Gamma \vdash P$, Δ and $\Gamma, Q \vdash \Delta$. (This is the 'classical' rule for introducing an implication as a premise, in contrast to the 'intuitionistic' rule of III.2.)

(iv) Derive, $\forall X N_1, \Gamma \vdash \Delta$ from $N_1[X/C]$, $\forall X N_1, \Gamma \vdash \Delta$, where X is the second order variable in N, introduced in I.2.1, N being of the form $\forall X N_1$ and $N_1[X/C]$ is obtained from N_1 by substituting the formula C for the variable X (corresponding to the comprehension axiom $\exists X \forall x [X(x) \leftrightarrow C]$).

---

(6) This implies $(\forall A) CF(A)$; actually in $\mathscr{PRA}$ from (the assumption that all Σ^0_1 sentences derivable in $\mathscr{UA}$ are true, that is, from) the reflection principle for $\mathscr{UA}$ applied to Σ^0_1 sentences. This provides an alternative proof of the result quoted above from Technical Note II of [13].

We shall apply (iv) only to formulae C which do *not* contain the variable x. In fact, renaming the bound variables in I.2.1, N_1 stands for

$$\{X(0)\wedge\forall u\forall v(u=v\to[X(u)\leftrightarrow X(v)])\wedge\forall w[X(w)\to X(sw)]\}\to X(x)$$

and so $N_1[X/C]$ stands for

$$\{C\wedge\forall u\forall v[u=v\to(C\leftrightarrow C)]\wedge\forall w(C\to C)\}\to C.$$

Abbreviation. Let C_1 stand for $\forall u\forall v[u=v\to(C\leftrightarrow C)]$ and C_2 for $\forall w(C\to C)$. Both C_1 and C_2 are (logically) derivable.

Suppose then we have a derivation, of A in $\mathscr{UA}$, which we call d_A if A is of the form $\forall nA'$ and d_B if A is of the form $\neg\exists nB'$. Since $\mathscr{UA}$ is (tacitly assumed to be) closed under all familiar rules we have derivations of

$$N\vdash A' \quad \text{resp. } N\wedge B'\vdash\emptyset$$

say d'_A, resp. d'_B. (This passage can be proved, in $\mathscr{PRA}$, also for $\mathscr{CFA}$, but needs a little argument familiar from 'inversion' theorems.)

We assume without loss of generality that the variable x (in the end formula) occurs *nowhere bound* in d'_A or d'_B. This is why we renamed the bound variables in N_1. Consequently, we may add the premise $\forall XN_1$, whose sole free variable is x, to the *LHS* of each sequent occurring in d'_A or d'_B without affecting any side condition restricting quantification rules. Let d^+_A and d^+_B be the derivations, of $\mathscr{UA}$, obtained by this 'weakening'. (It is here tacitly assumed that each rule of $\mathscr{UA}$ is monotone, that is, additional formulae may be added both on the *LHS* and the *RHS*.)

To eliminate cuts, by 'absorption' in $\forall XN_1$, consider any subderivation of d^+_A or d^+_B of the form

$$\begin{array}{cc} d_1 & d_2 \\ \forall XN_1, \Gamma_1\vdash C & \forall XN_1, \Gamma_2, C\vdash\Delta' \\ \multicolumn{2}{c}{\forall XN_1, \Gamma_1, \Gamma_2\vdash\Delta'} \end{array}$$

Our aim is to replace this last cut by a suitable application of (iii) to derive

$$N_1[X/C], \forall XN_1, \Gamma_1, \Gamma_2\vdash\Delta'$$

followed by (iv), which absorbs $N_1[X/C]$.

We shall apply (iii), with $\Gamma=\Gamma_1\cup\Gamma_2\cup\{\forall XN_1\}$, $Q=C$, $P=C\wedge\wedge C_1\wedge C_2$, where C_1 and C_2 were defined just after the explanation of the rule (iv), and $\Delta=\Delta'$. We 'weaken' d_1 by adding Γ_2 on the *LHS* and Δ' on the *RHS*, thus getting a derivation, say, d_3 of $\forall XN_1, \Gamma_1, \Gamma_2\vdash C, \Delta'$. Similarly, we weaken d_2 by adding Γ_1 throughout on the *LHS*, thus getting a derivation, say d_4, of $\forall XN_1, \Gamma_1, \Gamma_2, C\vdash\Delta'$.

Also we have derivations, say d_3' and d_3'' of $\vdash C_1,\ \Delta'$ resp. $\vdash C_2,\ \Delta'$.

$$\dfrac{\dfrac{d_3'}{\forall X N_1,\ \Gamma_1,\ \Gamma_2 \vdash C,\ \Delta'} \qquad \dfrac{d_3'' \qquad d_3''}{\vdash C_1,\ \Delta' \qquad \vdash C_2,\ \Delta'}}{\vdash C_1 \wedge C_2,\ \Delta'} \qquad \dfrac{d_4}{\forall X N_1,\ \Gamma_1,\ \Gamma_2,\ C \vdash \Delta'}$$

$$\forall X N_1,\ \Gamma_1,\ \Gamma_2 \vdash C \wedge C_1 \wedge C_2,\ \Delta'$$

As desired we get a derivation of $\forall X N_1,\ N_1[X/C],\ \Gamma_1,\ \Gamma_2 \vdash \Delta'$ by means of (iii) (and without the cut above) since $N_1[X/C]$ is $(C \wedge C_1 \wedge C_2) \to C$.

Lastly we absorb $N_1[X/C]$ by rule (iv).

Each cut of d_A, resp. d_B can be replaced in turn (and arbitrary order) to yield derivations in $\mathscr{CFA}$ of

$$N \vdash A' \quad \text{resp.} \quad N \wedge B' \vdash \emptyset.$$

Applying now the (cut free) rule (i), resp. (ii) we get

$\vdash (N \to A')$ and hence $\vdash \forall n A'$, resp. $\exists x (N \wedge B') \vdash \emptyset$ and hence $\vdash \neg \exists n B'$

Remarks. (a) To avoid surprises the reader should consider the case of a closed false Σ_1^0 formula A' for which $CF(A')$ cannot be proved in $\mathscr{PRA}$, in fact not even in $\mathscr{UA}$; while, by the example, the addition of the dummy quantifier $\forall n$ changes the logical status: $\vdash_{\mathscr{PRA}} CF(\forall n A')$. This is quite consistent with the idea that the impredicative premise N, in $N \to A'$, introduces totally new ways of proving $\forall n A'$ even for closed A'; cf. the, admittedly extreme, case of $\forall x\ (x \neq x \to A')$.

(b) For the sake of the application in I.2.8.2 we note the following variant of the *example* above.

For each A of the form $\neg \forall n \exists m C'$ we have $\vdash_{\mathscr{PRA}} CF(A)$.

Proof. Suppose then we have a derivation d in $\mathscr{UA}$ of A, that is, of

$$(*) \qquad \forall x [N \to \exists y (N' \wedge C')] \vdash \emptyset,$$

where N' is $N[x/y]$. To obtain a derivation in $\mathscr{CFA}$ it is sufficient to derive

$$(**) \qquad N[x/0] \to \exists y (N' \wedge C'[x/0]), \qquad \forall x [N \to \exists y (N' \wedge C')] \vdash \emptyset$$

from which (*) follows by the (cut free) rule: derive $\forall x B \vdash \emptyset$ from $B[x/t]$, $\forall x B \vdash \emptyset$.

Finally (**) follows by rule (iii) from

$$A \vdash N[x/0] \quad \text{and} \quad A,\ \exists y (N' \wedge C'[x/0]) \vdash \emptyset$$

by taking $\Gamma_1 = \emptyset$, $\Gamma_2 = \{A\}$, $\Delta = \emptyset$, $P = N[x/0]$ and $Q = \exists y (N' \wedge \wedge C'[x/0])$. But $\vdash N[x/0]$ is derivable outright in $\mathscr{CFA}$ and a derivation d'

in $\mathscr{UA}$ of A, $\exists y(N' \wedge C'[x/0]) \vdash \emptyset$ can be converted into one of $\mathscr{CFA}$ as in the previous example with $B' = C'[x/0]$. Finally we note that d' can be obtained from d by monotonicity.

(c) Both for formulae A of the *example* and of the *variant* in (b) above we could have referred to very general syntactic conditions, given by Girard [4], which ensure $\vdash_{\mathscr{CFA}} CF(A)$. For our present applications this result would have been enough; our sharpening, that is, the restriction of metamathematical methods to $\mathscr{PRA}$, is not needed. The significance of the sharpening lies elsewhere, as discussed in III.2. — However, pedagogically, our proofs have permanent value: If one is principally interested in *literal* Gödel sentences, of $\mathscr{CFA}$ or $\mathscr{CFA}_R$, our example and its variant are enough (for the applications in I.2.6 and I.2.8.2), and their proofs are shorter than that of Girard's general theorem.

3. Demonstrable instances of deductive equivalence and of the fundamental conjecture. As already noted in the introduction

$$\{[\vdash^-(A \leftrightarrow B)] \wedge CF(A) \wedge CF(B)\} \rightarrow (A \equiv B)$$

can be proved in $\mathscr{PRA}$. For if $\vdash^-(A \leftrightarrow B)$, then $\vdash^+(A \leftrightarrow B)$ and so, since $\mathscr{UA}$ is demonstrably closed under *modus ponens* also $\vdash_{\mathscr{PRA}} [(\vdash^+ A) \leftrightarrow (\vdash^+ B)]$. Now we apply $CF(A)$ and $CF(B)$.

3.1. Though $\vdash^- [CF(A) \wedge CF(B)]$ is certainly a sufficient condition for

$$(*) \qquad \vdash^- [\vdash^-(A \leftrightarrow B) \rightarrow \vdash^-(A \equiv B)],$$

it is evidently not necessary (take a false Σ^0_1 formula A and $B = A$). More importantly, for our present applications, for example in I.2.8.1, we (have and) use $(*)$ when A is the negation of a Π^0_1 formula and B is a (false) Σ^0_1 formulae. Specifically

A is of the form $\neg \forall x(N \rightarrow P)$ and B is $\exists x(N \wedge \neg P)$.

The reader may like to give a direct argument, similar to that of 1.a, for

$$\vdash_{\mathscr{PRA}} [(\vdash^- A) \rightarrow (\vdash^- B)].$$

(The proof of $\vdash_{\mathscr{PRA}} [(\vdash^- B) \rightarrow (\vdash^- A)]$ is even simpler.) In any case, Girard [4] gives rather general syntactic conditions on A and B which ensure $(*)$ and cover the particular A and B above.

Discussion. It is not satisfactory to have to check $\vdash^-(A \equiv B)$ by *ad hoc* arguments as above. Put differently, we'd like to be able to say what $\equiv$ 'means' in familiar terms. For example, to return to the remark

in I.2.8.1, let Con^- and, say, Con_1^- be the two formulations of the consistency of $\mathscr{CFA}$ given by

$$\forall A\,[(\vdash^- A)\to\neg\vdash^-(\neg A)] \quad \text{(as before) and} \quad \neg\exists A\,[(\vdash^- A)\wedge(\vdash^-\neg A)].$$

Evidently $\vdash^-(\mathrm{Con}^-\leftrightarrow\mathrm{Con}_1^-)$; but, by above, also $\vdash^-(\mathrm{Con}^-\equiv\mathrm{Con}_1^-)$. Should we expect this because Con^- and Con_1^- are 'synonymous'?

Or, again, let the Π_1^0 formula U express the Riemann hypothesis and suppose U is independent of $\mathscr{CFA}$. Then neither $0=0\equiv U$ nor $0=1\equiv U$ (nor $0=0\equiv\neg U$ nor $0=1\equiv\neg U$) can be derived in $\mathscr{CFA}$ (since $[\vdash^-(\neg U)]\leftrightarrow\neg U$ can be). If $\vdash^- U$ then, of course $\vdash^-(0=0\equiv U)$ also $\vdash^-[0=1\equiv(\neg U)]$; but if $\vdash^-(\neg U)$, then $\vdash^-(0=1\equiv U)$ is false. Evidently this expresses that the possibilities of deriving $\neg U$ in $\mathscr{CFA}$ are 'similar' to those of deriving $0=1$ (essentially by computation), but the possibilities of deriving U are of a different 'kind'. Is this connected with synonymity? These and similar questions are discussed in III.3.

III. DISCUSSION OF SOME GENERAL ISSUES RAISED IN THE INTRODUCTION

We shall use the notation of the introduction.

Being new the results of Parts I and II, trivially, constitute 'progress' over earlier work. Also — if one is at all interested in self-referential propositions and/or $\mathscr{CFA}$ — there can hardly be any more *natural* questions than those we have chosen:

What have the different formulae which express their own derivability, resp. underivability, in common?

How, if at all, do the metamathematical means needed to establish the normal form theorem $CF(A)$ for (all derivations in $\mathscr{UA}$ of) a given formula A depend on A?

We now want to see whether, beyond its natural appeal, the work has interest for more general logical issues; in modern jargon, we want to analyze its *relevance*. Of course, the overriding question concerns the relevance of $\mathscr{CFA}$ itself before one even considers that of questions *about* $\mathscr{CFA}$ (particularly since, at least so far, $\mathscr{CFA}$ has not been very useful as a tool, for example, in the study of $\mathscr{UA}$).

1. Hilbert's programme. We assume the elementary exposition in [15]; App. II B. Neither Jeroslow's general extension [9] of Gödel's second theorem (which, by II.1a, applies immediately to $\mathscr{CFA}$) nor our extension of the sharper form: $A_G \leftrightarrow \mathrm{Con}_{\mathscr{CFA}}$, is relevant, simply because the usual conditions on systems, listed in [18] or A.2, are necessary if a formalization of mathematical reasoning is to be *adequate* for Hilbert's programme. And for such systems (even the sharper form of) Gödel's theorem is classical. Let us spell out the two adequacy conditions on a system $\mathscr{F}$:

(a) Demonstrable completeness w.r.t. Σ^0_1 formulae is needed to assure us that elementary mathematics (with a constructive existential quantifier) can be reproduced in $\mathscr{F}$ at all. This condition is satisfied by $\mathscr{CFA}$.

(b) Demonstrable closure under cut (and in the quantifier free case also under substitution) is also needed (7) because cut is constantly used in mathematics. Realistically speaking, a (meta)mathematical *proof* of such closure is needed and not a case study of mathematical texts because cut — like most logical inferences — is often used without being mentioned; in contrast, for example, to the use of mathema-

tical axioms. This closure condition is *not* satisfied by $\mathscr{CFA}$, by I.1.3.

Incidentally, we do not see any relevance to Hilbert's programme of the additional information (over Gödel's theorem) supplied by Löb's [18] though, as mentioned in the introduction, Löb's includes Gödel's for the usual (adequate) systems (7).

Having made all this clear, in particular the inadequacy of $\mathscr{CFA}$ for Hilbert's programme, we must now emphasize this: at the present time, 40 years after [5], *it would be simply absurd to judge proof-theoretic work mainly, let alone solely, by its relevance to Hilbert's original programme, because this programme is itself of dubious relevance.* One of us would go so far as to say that its principal epistemological value was this: it provided the concepts in terms of which more or less natural assumptions about mathematical reasoning could be formulated precisely enough to be put in their place. The assumptions are *verified for the bulk of mathematical practice* in as much as this practice can be formalized in remarkably weak (sub)systems (of set theory). The assumptions are *refuted in principle* since they *do not apply to all possible valid mathematical reasoning.* Also — and, for one of us, this is philosophically by far the most significant result of work on Hilbert's programme — the *original epistemological claims for the programme*, concerning its relevance to certainty, *have not been established.* As a matter of empirical fact, our confidence in a part of practice was *not* increased, *when abstract* (set theoretic) *concepts were successfully eliminated*; certainly much less than by a more precise analysis of *which* sets we are talking about (which segments of the cumulative hierarchy). This fact of mathematical experience seems — again, to one of us — just as convincing evidence against those assumptions about mathematical reasoning which are behind Hilbert's programme as its theoretical refutation mentioned above. Both of us agree that the programme and, in particular, its distinction between finitist and non-finitist reasoning are natural, especially when one *begins* to reflect on mathematics; but we cannot agree on the objective epistemological significance of this distinction.

On the other hand neither of us doubts the permanent value of proof theory if this theory is separated from those philosophical doubts (about the validity of currently used principles) which provided the original *raison d'être* for Hilbert's programme. What we envisage is an analysis

---

(7) To avoid misunderstanding, both Jeroslow's and Löb's improvements are essential for the deeper mathematical study of formal systems; as to Löb's, see e.g. [11], p. 167, 3.3421. Jeroslow's improvement shows that the extension of Gödel's original theorem to $\mathscr{CFA}$ needs merely completeness of $\mathscr{CFA}$ for Σ^0_1 formulae; only the extension of the sharpened form: $A_G \to \mathrm{Con}_{\mathscr{CFA}}$ uses more recondite properties of $\mathscr{CFA}$.

of the *structure of proofs*, providing concepts in terms of which we can state facts (about proofs) which we really want to know.

2. $\mathscr{CFA}$ and the structure of proofs in analysis. One of us [22] introduced $\mathscr{CFA}$ to replace the essentially epistemological consistency problem (for $\mathscr{UA}$) — which is open only if the metamathematical methods are restricted — by the mathematical normal form problem (or 'fundamental conjecture') — which was open whatever (valid) methods were allowed. The strategy behind this step was this.

As we should put it now, since say $\neg\vdash_{\mathscr{CFA}} 0=1$ is derivable in $\mathscr{PRA}$, $\mathrm{Con}_{\mathscr{UA}}$ (or, equivalently, by I.2.5b, $\mathrm{Con}_{\mathscr{CFA}}$) follows, still in $\mathscr{PRA}$, from $CF(0=1)$. So there was a chance that a proof by the 'light of nature' of $CF(A)$ for all A or even of $CF(0=1)$ would somehow lead us to the discovery of metamathematically relevant methods, besides providing the intrinsically interesting fact that $CF(A)$ is *true* for all A. (Of course $CF(0=1)$ is true since $0=1$ is not derivable in $\mathscr{UA}$.) This hope was quit econsistent with the fact that, by Gödel's theorem, $\forall A[CF(A)]$ cannot be proved in $\mathscr{UA}$, since $CF(0=1)\rightarrow\mathrm{Con}_{\mathscr{UA}}$ in $\mathscr{PRA}$. The hope was that the methods used would be *incomparable* with those of $\mathscr{UA}$ in the way in which Gentzen's use of ε_0-induction applied to *quantifier-free* predicates is incomparable with first order arithmetic.

It is fair to say that this hope was justified for certain subsystems of $\mathscr{UA}$ [23]. But all known proofs of $\forall A[CF(A)]$ for the whole of $\mathscr{UA}$ use methods whose (natural) formulation simply *includes* the whole of $\mathscr{UA}$. More precisely, this applies to model theoretic proofs; proofs via so-called normalization [4] make *full* use of impredicative comprehension principles (only the logical rules may now be replaced by formally intuitionistic ones). In short, the hope mentioned has not been fulfilled for the whole of $\mathscr{UA}$.

Before proceeding further, it is proper to go into terminology, the use of 'cut free', because — as so often — the choice of terminology expresses one's views of what is basic. If the topic of self-referential propositions treated in Part I were a principal preoccupation, the terminology would indeed be *justified by the general theorems quoted in Part* I! Demonstrable closure under cut is the only metamathematical property of the usual systems actually needed to prove the results (about their self-referential propositions) which distinguish those systems from $\mathscr{CFA}$. In contrast, other expositions of — what are called — cut free systems, for example in [12], p. 329, emphasized the so-called *subformula property*; it so happens that all the inference rules of familiar systems possess this property except for the cut rule — at least, if one overlooks progressive abuse of the ordinary meaning of 'subformula'. More precisely, in propositional logic, a subformula of A is, literally, a part of A; in first order

predicate logic with function symbols each of the infinitely many formulae $A[x/t]$, where t is an arbitrarily complicated function term, is a subformula of $\exists x A$ or $\forall x A$; in second order logic each $A(X/T)$ is a subformula of $\exists X A$, where T is an arbitrary abstraction term, which may contain $\exists X A$ itself as a proper part! (However, the transitive closure of the relation: being an immediate proper subformula of, is wellfounded.) We recognize that, with this widening of the meaning of 'subformula', progessively more imagination is needed to make use of the subformula property; and so there are reasonable doubts about the proper choice of meaning. There is a need for discovering tests for choosing between alternatives, and only research will show how far one can go with cut free systems.

Speaking for ourselves — and, perhaps, not for the majority of proof theorists — we are persuaded, by what we know of $\mathscr{CFA}$ and related systems, that they have independent interest; as *principal* objects of study for a structural theory of proofs (and not merely as an auxiliary towards a consistency proof for $\mathscr{UA}$ for which purpose $\mathscr{CFA}$ was originally introduced). Our confidence is not at all shaken by the *inadequacy* of $\mathscr{CFA}$, observed in Section 1, for the representation of actual mathematical reasoning. On the contrary, since mathematicians *use* proofs without making them (and their structure) into *objects* of study, we should expect a *conflict* between (i) the aims of Hilbert's programme and (ii) the aims here considered. The faithful representation of actual reasoning needed for (i) contains of course all the shortcuts used in mathematics, general lemmas $\forall x P$ of which only a particular case is used (by combining instantiation: $\forall x P \to P[x/t]$ and *modus ponens*, as in $\mathscr{UA}$ but not $\mathscr{CFA}$). In contrast, for (ii) one will want to analyze explicitly, for example, what part of the proof of a general lemma is actually relevant to the particular case needed for the conclusion; and at least for Σ^0_1 sentences, $\mathscr{CFA}$ automatically excludes steps which are not directly relevant to the (Σ^0_1) theorem derived. — To avoid misunderstanding; we do not claim that the *exact choice of rules* of $\mathscr{CFA}$ (which was of course not important for the purpose for which $\mathscr{CFA}$ was originally introduced) is *exactly* right for (ii); we should only claim that *something like* $\mathscr{CFA}$ is needed.

From the metamathematics to the mathematics of the normal form theorem: a fresh start. We recall that (something like) $\mathscr{CFA}$ provides a more explicit analysis of proofs (than the usual systems like $\mathscr{UA}$) and that the normal form theorem is true, that is

$$\forall d \forall A [\mathrm{Prov}_{\mathscr{UA}}(d, \ulcorner A \urcorner) \to \exists d' \mathrm{Prov}_{\mathscr{CFA}}(d', \ulcorner A \urcorner)].$$

Remark. More precisely, in the terminology of the next section, we should say

$$\forall d \forall A [\mathrm{Prov}_{\mathscr{UA}}(d, \ulcorner A \urcorner) \to \exists d' \exists A' \{\mathrm{Prov}_{\mathscr{CFA}}(d', \ulcorner A' \urcorner) \wedge (A \,\mathrm{syn}\, A')\}]$$

(where *syn* means: synonymous to), that is

To each proof ($\bar{d}$ expressed by the formal derivation) d in $\mathscr{U}\mathscr{A}$ there is *some* proof d' in $\mathscr{C}\mathscr{F}\mathscr{A}$ of the same (proposition, expressed by the same or a synonymous) end formula.

We now ask for more:

To associate a *particular* d' with d (from among all the

$$\{d' \colon \exists A' \mathrm{Prov}_{\mathscr{C}\mathscr{F}\mathscr{A}}(d', \ulcorner A' \urcorner) \wedge A \,\mathrm{syn}\, A'\}$$

such that d' provides a meaningful analysis of d, for example, d' exhibits the part of d actually relevant to the conclusion A ([8]).

NB. *Every* (metamathematical) proof of the normal form theorem provides some mapping: $d \mapsto d'$; if none other than the 'trivial' one of looking for the first d' in some ω-ordering of derivations in $\mathscr{C}\mathscr{F}\mathscr{A}$, with the same end formula as d. Our aim is much more ambitious: d' should have something to do with the *structure* of (the proof expressed by) d, not merely have the same end formula. — *Warning*: we are not primarily concerned with the metamathematical status of the mapping, say μ: $d \mapsto d'$, that is, with how 'elementary' the definition principles used for introducing μ are, or how 'elementary' the proof of

$$\mathrm{Prov}_{\mathscr{U}\mathscr{A}}(d, \ulcorner A \urcorner) \rightarrow \mathrm{Prov}_{\mathscr{C}\mathscr{F}\mathscr{A}}(\mu d, \ulcorner A \urcorner)$$

is. The principal aim is not metamathematical but mathematical: *What structure is preserved by* μ? It is a separate question whether different proofs of the normal form theorem, by more or less elementary metamathematical principles, *turn out* to provide mappings μ which preserve (significantly) different structure. To be specific, we consider two mappings.

The first, and best known, was introduced by Prawitz for systems related to ours, so-called systems of natural deduction. When the literature discusses *normalization theorems* it is tacitly assumed that Prawitz' particular sequences of normalization steps are applied to provide a mapping, which we shall call λ; λ for logical because this normalization procedure depends only on the *logical form* of the rules used (in building up the derivation d), not on the *content* of the propositions proved. (Prawitz' method is uniform.) — As a kind of pun, there is also another reason for using λ (logical) because Prawitz' own reason for his choice of normalization steps was originally dictated by a certain *operational semantics of the logical particles*. Though we find this semantics far removed from the ordinary intended meaning (constructive or non-constructive), we are inclined to believe that the formal relation: $d \overset{\lambda}{\mapsto} d'$ has significance also for less

([8]) In [14] one of us considered the possibility that the proofs ($\bar{d}$ expressed by) d and ($\bar{d}'$ expressed by) d' are identical. This possibility is even more implausible than appears from [14], as indeed many people observed. The descriptions under (i) and (ii) in the last paragraph were chosen to make this clear. However, the application made in [14], § 1(b) of the particular mapping $d \mapsto d'$ is not affected.

dubious applications (than those to operational semantics); this view is in accordance with experience in science generally where facts and methods discovered in pursuing a false theory are often useful for a later sound theory.

We now wish to consider *different mappings* suggested by our discovery that the normal form theorem $CF(A)$ can be proved by obviously different methods for certain classes of A. Inasmuch as the mappings depend on A, we have here normalization procedures according to *content*. We choose a class $\boldsymbol{A}$ of formulae such that

$$A \in \boldsymbol{A} \rightarrow [CF(A) \text{ can be proved in } \mathscr{PRA}],$$

but the termination of Prawitz' procedure for $A \in \boldsymbol{A}$ cannot be proved in analysis! This choice is made to highlight the difference in familiar terms; as we said above the significance of the difference between the two procedures lies, for us, in the different structure (of d) preserved by them.

EXAMPLE (chosen to apply both to $\mathscr{UA}$ and to its formally intuitionistic version). $\boldsymbol{A}$ consists of formulae of the form

$$\forall p(p \rightarrow p) \rightarrow B,$$

where B is arbitrary and p is a propositional variable. (For $\mathscr{UA}$ itself the class of $\forall p(p \vee \neg p) \rightarrow B$ would do equally well.)

(a) There is a procedure which can be shown in $\mathscr{PRA}$ to convert every derivation d (in $\mathscr{UA}$) of A into one in $\mathscr{CFA}$.

We can use the premise $\forall p(p \rightarrow p)$ to 'absorb' all cuts. First, without spoiling the proof structure, we add the premise $\forall p(p \rightarrow p)$ to each left sequent in the given derivation d to form, say, d_0. Suppose then a subderivation of d_0 has the form

$$\dfrac{\begin{array}{cc} d_1 & d_2 \\ \forall p(p \rightarrow p), \Gamma \vdash C & \forall p(p \rightarrow p), \Gamma', C \vdash C' \end{array}}{\forall p(p \rightarrow p), \Gamma, \Gamma' \vdash C'} \text{ cut}$$

We recall the general (intuitionistic) *cut free* rule for 'introducing' an implication on the left, namely

$$\dfrac{\Gamma_1, A \rightarrow B \vdash A \qquad \Gamma_2, B \vdash D}{\Gamma_1, \Gamma_2, A \rightarrow B \vdash D}$$

and the — also cut free — absorption rule for 'introducing' a universal quantifier (on the left), namely

$$\dfrac{C \rightarrow C, \forall p(p \rightarrow p), \Gamma_3 \vdash D}{\forall p(p \rightarrow p), \Gamma_3 \vdash D}.$$

Next, again without spoiling the proof structure we 'weaken' the whole of d_1, by adding $C \to C$ to each sequent on the left to form, say, d_3. Consider now

$$\begin{array}{cc} d_3 & d_2 \\ \forall p(p \to p),\ C \to C,\ \Gamma \vdash C & \forall p(p \to p),\ \Gamma',\ C \vdash C' \end{array}$$

We now replace the cut by application of the two cut free inferences, where A and B are replaced by C; D by C';

Γ_1 by $\forall p(p \to p), \Gamma$; $\quad \Gamma_2$ by $\forall p(p \to p), \Gamma'$; $\quad$ and $\quad \Gamma_3$ by Γ, Γ'

Evidently, if there are n cuts, this procedure reduces d to a cut free derivation of $\forall p(p \to p) \to B$ in n steps.

(b) Let B be a false Σ_0^0 (or Σ_1^0) formula. Then it is not possible to prove by metamathematical methods formalized in $\mathscr{U}\mathscr{A}$ that all derivations d of $\forall p(p \to p) \to B$ can be normalized by means of Prawitz' procedure.

To see this consider those derivations d (if any) which have the form

$$\begin{array}{l} d' \\ B \\ \forall p(p \to p) \to B \quad \text{(weakening)}, \end{array}$$

where we assume without loss of generality that neither d' nor B contains the propositional variable p. When we apply Prawitz normalization procedure, the part

$$\begin{array}{l} d' \\ B \end{array}$$

of d is treated independently of the last line; contrary to the procedure in (a) one does not introduce the formula $\forall p(p \to p)$ into d'. Thus if all our d are normalized by Prawitz' procedure then so are all derivations d' of B. But this can certainly not be proved in $\mathscr{U}\mathscr{A}$; if it could, that is, if

$$\vdash_{\mathscr{U}\mathscr{A}} [(\vdash_{\mathscr{U}\mathscr{A}} B) \to (\vdash_{\mathscr{C}\mathscr{F}\mathscr{A}} B)]$$

we should certainly have

$$\vdash_{\mathscr{U}\mathscr{A}} [(\vdash_{\mathscr{U}\mathscr{A}} B) \to B],$$

since the reflection principle for *all* Σ_0^0 (or Σ_1^0) formulae of $\mathscr{C}\mathscr{F}\mathscr{A}$ can be proved in $\mathscr{P}\mathscr{R}\mathscr{A}$, that is

$$\vdash_{\mathscr{P}\mathscr{R}\mathscr{A}} [(\vdash_{\mathscr{C}\mathscr{F}\mathscr{A}} B) \to B].$$

But then, by [18], $\vdash_{\mathscr{A}} \sim B$ which is impossible since B is false.

If B is true (and therefore $\vdash_{\mathscr{CFA}} B$), each derivation d of $\forall p(p\to p)\to B$ can of course be normalized both by Prawitz' procedure and that of (a); but the cut free derivations obtained are different (unless the derivation is cut free to start with).

Remarks. (i) Not all the metamathematics of Prawitz' λ goes beyond $\mathscr{PRA}$. For example, let d_A^* be the usual derivation in $\mathscr{CFA}$ of $A\to A_p$, where A_p is a prenex normal form of A. For variable d (over derivations in $\mathscr{CFA}$) let $\pi_A d$ be (the derivation) d followed by cut with d_A^*. Then we easily prove in $\mathscr{PRA}$ that Prawitz' normalization procedure applied to $\pi_A d$ terminates, that is, his λ restricted to $\pi_A d$ is primitive recursive.

(ii) In our early work, before Girard's general result, we established $CF(A)$ in $\mathscr{PRA}$ for many classes of formulae other than A of the example above. For the purpose of our result on Gödel sentences of $\mathscr{CFA}$ there is no gain whatever in the 'reduction' to $\mathscr{PRA}$; indeed, as Girard observes, his result can be established in quite weak subsystems of analysis (and this too presents no gain, for our particular result, as long as we stay *within* analysis). Certainly, epistemologically, the replacement of Girard's metamathematical principles by $\mathscr{PRA}$ is pointless. (Probably also the *lengths* of the procedures arising from Girard's proof and ours don't differ too much.) But the resulting mappings are different and, it would seem, preserve quite different structure, more subtle than mere length.

(iii) Though apparently merely cute, the following question seems interesting (to one of us): Is there a normalization procedure or mapping: $d\mapsto d'$ specially adapted to derivations of Gödel sentences? as the mapping in (a) of the example was adapted to derivations of formulae in A. — The question seems merely cute in as much as all Gödel sentences are underivable and the 'specially adapted' procedure could never be applied.

3. Henkin's problem [6] and the relation of synonymity (between formulae expressing the same proposition). What more do we know when we know that two expressions are synonymous than if we merely know that they are formally equivalent (in some familiar system)? or, more generally:

What more do we know when we have 'intensional' equality and not only extensional equivalence in all models (of a familiar class of models)?

The questions are not merely rhetorical. We know from ordinary mathematical experience the *surprising* discovery how often our assertions depend only on some simple extensional features (such as the *graph* of a function) even if we think of the objects intensionally (such as rules for computing a function). And, at least for the usual systems $\mathscr{F}$ satisfying the conditions listed by Löb [18], we have the surprising discovery of equivalence between all Gödel sentences, resp. all Henkin sentences when

we merely assume formal equivalence between A_G and $\neg\vdash_{\mathscr{F}} A_G$ and between A_H and $\vdash_{\mathscr{F}} A_H$ (and not synonymity).

Of course these discoveries are satisfactory if one wants to *avoid* an analysis of the more delicate relation of synonymity. Also they suggest that such an analysis may well have only limited use. But above all these discoveries show that one will have to *look* for areas where the relation of synonymity is genuinely significant; otherwise one is bound to overdramatize some puzzles about synonymity into 'problems', familiar from traditional philosophy. — We are quite struck by the fact that, in the case of Henkin sentences for $\mathscr{CFA}$, we were able to get obviously stronger conclusions for *literal* A_H (where A_H and $\vdash_{\mathscr{CFA}} A_H$ are synonymuos). Also we find encouraging the fact that $\mathscr{CFA}$ (which seems to us more suited for logical *analysis* than the usual systems) exhibits this phenomenon — in contrast to $\mathscr{UA}$. It is more difficult to be specific about a *theory of synonymity*. We shall confine ourselves to quite elementary examples and generalities.

EXAMPLE.

(a) Is the relation $A \equiv\!\!\!\equiv B$ a promising approximation to A syn B, where $A \equiv\!\!\!\equiv B$ is the conjunction of

$$\vdash^{-}(A \leftrightarrow B) \text{ and } \vdash^{-}[(\vdash^{-} A) \leftrightarrow (\vdash^{-} B)],$$

that is, of formal and demonstrable deductive equivalence in $\mathscr{CFA}$ — or, again, $[\vdash_{\mathscr{CFA}}(A \leftrightarrow B)] \wedge [\vdash_{\mathscr{CFA}}(A \equiv B)]$ in the notation of I.1.1? The answer is *No* on the assumption that *negation preserves synonymity*. (Take for A the formula $\forall p(p \rightarrow p)$ and $0 = 0$ for B. Evidently, $A \equiv\!\!\!\equiv B$ and $\vdash^{-}[(\neg A) \leftrightarrow (\neg B)]$. But $\neg A \equiv\!\!\!\equiv \neg B$ is false since $\vdash^{-}(\neg\vdash^{-}\neg B)$ (B being numerical), but not $\vdash^{-}\neg\vdash^{-}\neg A$. It is sufficient to establish $CF(\neg A)$ since $\neg\vdash^{+}\neg A$ is not even derivable in $\mathscr{UA}$ itself.

(Hint. as in the analysis in III.2 above, the last step in a cut free derivation of $\neg A$ infers $\neg A$ from

$$\forall p(p \rightarrow p) \vdash \emptyset .$$

The formula $\forall p(p \rightarrow p)$ on the left can then be used to absorb all cuts.)

(b) Is the relation $A \equiv_3 B$, equivalence in all 3 valued models considered by Girard [4], a better approximation than $\equiv\!\!\!\equiv$? The answer is *Yes* since all logical operations do preserve $\equiv_3$ and also

$$A \text{ syn } B \Rightarrow A \equiv_3 B \Rightarrow A \equiv\!\!\!\equiv B .$$

Generalities. Despite the two satisfactory properties of $\equiv_3$ just mentioned, we see little evidence that this or any other formal relation provides

a *complete general* theory of synonymity; 'general' in the sense to which we have become accustomed: applicable to all expressions of such familiar logical languages as propositional or predicate logic. But had we not better revise our ideas? For which (useful) concepts — besides the singular case of *classical* validity of first order logic — do we have complete general theories? Certainly not for second order logic; and yet there are a great number of 'partial' results (for example in connection with so-called axioms of infinity), at least as interesting as the complete theory for first order logic. And even with a much more modest requirement of 'completeness' on languages, we have no complete theory of diophantine equations, but many powerful results in that subject. Perhaps synonymity requires similar piecemeal study for progress.

We are also aware of the common pessimism about analyzing synonymity convincingly by a relation which lies properly between syntactic equality (up to renaming of variables) and formal equivalence. This pessimism is surely justified if we compare *synonymity* with such logical notions as validity or mechanical computability which possess an analysis which is, at least *post hoc*, almost immediately convincing; or, indeed, if we compare synonymity with the most familiar elementary geometrical or dynamical notions. Our own optimism derives from the discovery of such 'advanced' notions as that of *topological equivalence* (between geometric figures) which are used to analyze our spatial experience. Was there not a time when people were pessimistic about discovering a significant geometric relation properly between, say, metric congruence and mere 1–1 correspondence (introduced by Cantor)?

Appendix. Addenda to the literature

1. Jeroslow's variant of literal Gödel sentences ([9]; cf. the remark in I.1.6). Jeroslow considers, in place of literal Gödel sentences A_G, Σ_1^0 sentences A_J *of the form* $\vdash_{\mathscr{F}} \neg A_J$. Thus $\neg A_J$ is a literal Gödel sentence for $\mathscr{F}$ as defined in I.1.6.

Since A_J is a Σ_1^0 formula we get, in place of (*) in I.1.6,

$$\vdash_{\mathscr{F}} [\mathrm{Con}_{\mathscr{F}}(A_J) \rightarrow \neg A_J];$$

for $\vdash_{\mathscr{F}} (A_J \rightarrow \vdash_{\mathscr{F}} A_J)$ by completeness for Σ_1^0 formulae and $\vdash_{\mathscr{F}} [A_J \rightarrow \vdash_{\mathscr{F}} (\neg A_J)]$ because $\neg A_J$ is (of the form) $\vdash_{\mathscr{F}} (\neg A_J)$. But also

$$\vdash_{\mathscr{F}} [\neg A_J \rightarrow \mathrm{Con}_{\mathscr{F}}(A_J)]$$

holds because $[\neg \vdash_{\mathscr{F}} (\neg A_J)] \rightarrow \mathrm{Con}_{\mathscr{F}}(A_J)$. In contrast, the converse to (*) does not seem to be generally derivable; we should need

$$\neg(\vdash_{\mathscr{F}} A_G) \rightarrow [\vdash_{\mathscr{F}} (\vdash_{\mathscr{F}} (\vdash_{\mathscr{F}} A_G) \rightarrow \neg \vdash_{\mathscr{F}} (\neg \vdash_{\mathscr{F}} A_G)].$$

Remark. Jeroslow's improvement can also be obtained for any literal Gödel sentence A_G of I.1.6. Since A_G has the form $\neg B_G$ and $B_G \in \Sigma_1^0$, we have $\vdash_{\mathscr{F}} (B_G \rightarrow \vdash_{\mathscr{F}} B_G)$ and of course $\vdash_{\mathscr{F}} [B_G \rightarrow \vdash_{\mathscr{F}} (\neg B_G)]$; hence

$$\vdash_{\mathscr{F}} [A_G \leftrightarrow \mathrm{Con}_{\mathscr{F}}(B_G)].$$

2. Löb's theorem. The conditions in [18] are a little too complicated. We shall use a slightly different construction and thereby reduce the requirement to:

I. $\vdash_{\mathscr{F}} A$ *is* Σ_1^0; *and derivable in* $\mathscr{F}$ *if and only if* A *is derivable in* $\mathscr{F}$ (*local definition of*: *derivability in* $\mathscr{F}$; *a canonical derivability definition is naturally* Σ_1^0).

II. $\mathscr{F}$ *is closed w.r.t. positive implicational logic and satisfies the deduction theorem* (*for closed formulae*).

III. $\mathscr{F}$ *is complete for* Σ_1^0 *formulae.*

IV. *For any two formulae* A_0 *and* B_0

$$[\vdash_{\mathscr{F}} (B_0 \rightarrow A_0)] \rightarrow [(\vdash_{\mathscr{F}} B_0) \rightarrow (\vdash_{\mathscr{F}} A_0)]$$

is derivable in $\mathscr{F}$.

Then $\vdash_{\mathscr{F}}[(\vdash_{\mathscr{F}}A_0)\to A_0]$ *is derivable in* $\mathscr{F}$ *only if* $\vdash_{\mathscr{F}}A_0$.

Proof. Suppose $(\vdash_{\mathscr{F}}A_0)\to A_0$ is derivable in $\mathscr{F}$. Construct (unlike [18]) B_0 which is literally of the form

$$\vdash_{\mathscr{F}}(B_0\to A_0), \quad \text{and hence } \Sigma_1^0.$$

By I and III, $B_0\to\vdash_{\mathscr{F}}B_0$ (in $\mathscr{F}$) and, since $B_0\to\vdash_{\mathscr{F}}(B_0\to A_0)$ by construction

$$B_0\to[(\vdash_{\mathscr{F}}B_0)\to(\vdash_{\mathscr{F}}A_0)] \quad \text{by IV.}$$

By II, $B_0\to\vdash_{\mathscr{F}}A_0$ and so by assumption, $B_0\to A_0$ (in $\mathscr{F}$).
By I, $\vdash_{\mathscr{F}}(B_0\to A_0)$ and so $(\vdash_{\mathscr{F}}B_0)$.
By IV, $(\vdash_{\mathscr{F}}B_0)\to(\vdash_{\mathscr{F}}A_0)$ which, with $\vdash_{\mathscr{F}}B_0$ and II, yields $\vdash_{\mathscr{F}}A_0$.
Evidently, by II, $(\vdash_{\mathscr{F}}A_0)\Rightarrow\vdash_{\mathscr{F}}[(\vdash_{\mathscr{F}}A_0)\to A_0]$.

Comparison with an earlier discussion of Henkin's problem. Before Löb's solution for usual systems, one of us described, in [10], two systems such that one has derivable, the other refutable Henkin sentences. Since the concept of *canonical representation* was not available, the exposition is defective (like the rest of the literature), referring to different 'representations' of the derivability predicate. We therefore give here a civilized description of those systems, modified according to a suggestion of Henkin loc. cit.

Let $\mathscr{F}$ be a usual system.

(i) Let the formula Q_1 be of the form $\ulcorner Q_1\urcorner = \ulcorner Q_1\urcorner\vee\vdash Q_1$, and define $\mathscr{F}_1$ as follows.

Q_1 is an axiom, and, in addition, any derivation of $\mathscr{F}$ is also a derivation of $\mathscr{F}_1$.

$\mathscr{F}_1$ is clearly a formal system and

$$\vdash_1 A \textit{ is } \ulcorner A\urcorner = \ulcorner Q_1\urcorner\vee\vdash A.$$

Q_1 is a literal Henkin sentence of $\mathscr{F}_1$ and evidently derivable in $\mathscr{F}_1$. Also $\mathscr{F}$ and $\mathscr{F}_1$ have, demonstrably, the same theorems since Q_1 is derivable in $\mathscr{F}$.

(ii) Let the formula Q_2 be of the form $\ulcorner Q_2\urcorner \neq \ulcorner Q_2\urcorner\wedge\vdash Q_2$, and define $\mathscr{F}_2$ as follows.

Any derivation of $\mathscr{F}$ is a derivation of $\mathscr{F}_2$ unless its end formula is Q_2.

$\mathscr{F}_2$ is clearly a formal system, and

$$\vdash_2 A \textit{ is } \ulcorner A\urcorner \neq \ulcorner Q_2\urcorner\wedge\vdash A.$$

Q_2 is evidently a literal Henkin sentence of $\mathscr{F}_2$ and *refutable* in $\mathscr{F}_2$. Also $\mathscr{F}$ and $\mathscr{F}_2$ have, in fact, the same theorems since $\ulcorner Q_2\urcorner \neq \ulcorner Q_2\urcorner$ is not derivable in $\mathscr{F}$; but this is not demonstrable in $\mathscr{F}$ nor, *a fortiori*, in $\mathscr{F}_2$

since, for a usual system $\mathscr{F}$, *no* underivability result (about $\mathscr{F}$) can be demonstrated in $\mathscr{F}$.

In [6] Henkin used, informally, the notion of 'standard formal system adequate for recursive number theory'. Clearly, if 'adequacy' refers only to the set of theorems, $\mathscr{F}_1$ and $\mathscr{F}_2$ are as adequate as $\mathscr{F}$! (and they are *formal* systems). If adequacy requires demonstrable closure under *modus ponens* (and uniform completeness for primitive recursive arithmetic), then $\mathscr{F}_1$ is, but $\mathscr{F}_2$ is not adequate. In fact, let A be refutable but distinct from Q_2; then we have $\vdash_2(A \to Q_2)$ but *not* $\vdash_2[(\vdash_2 A) \to \vdash_2 Q_2]$. Since $A \neq Q_2$ and $\vdash \neg\vdash_2 Q_2$ we should have $\vdash(\neg\vdash A)$ which is impossible for our (usual) system $\mathscr{F}$. Thus $\mathscr{F}_2$ fails to be (demonstrably) *locally* closed under *modus ponens*.

For the record it is, perhaps, worth observing that *not even all literal Henkin sentences of $\mathscr{F}_2$ are equivalent.*

Let Q_2' be also of the form $\ulcorner Q_2 \urcorner \neq \ulcorner Q_2 \urcorner \wedge \vdash Q_2'$, but distinct from Q_2. (There are such Q_2' by footnote ([3]), p. 7.) Since

$$\vdash(\ulcorner Q_2' \urcorner \neq \ulcorner Q_2 \urcorner)$$

we have

$$\vdash(Q_2' \leftrightarrow \vdash Q_2')$$

and so Q_2' is also a Henkin sentence of the usual system $\mathscr{F}$. Thus, by Löb's theorem, $\vdash Q_2'$. But since Q_2' is distinct from Q_2

$$\vdash_2 Q_2'.$$

Remark. In contrast — but in accordance with footnote ([2]) (p. 6) of [10] — all Gödel sentences of $\mathscr{F}_2$ are demonstrably equivalent to $\mathrm{Con}_{\mathscr{F}}$. Such a sentence, say G_2, satisfies $\vdash_2(G_2 \leftrightarrow \neg\vdash_2 G)$ or, more fully,

$$\vdash_2[G_2 \leftrightarrow (\ulcorner G_2 \urcorner = \ulcorner Q_2 \urcorner \vee \neg\vdash G_2)].$$

If $\ulcorner G_2 \urcorner \neq \ulcorner Q_2 \urcorner$, we have $\vdash_2(G_2 \leftrightarrow \neg\vdash G_2)$ and so G_2 is a Gödel sentence for (the usual system) $\mathscr{F}$, hence demonstrably equivalent to $\mathrm{Con}_{\mathscr{F}}$. Finally, Q_2 is *not* a Gödel sentence for $\mathscr{F}_2$ since Q_2 is of the form $\vdash_2 Q_2$, hence not equivalent to $\neg\vdash_2 Q_2$.

3. Rosser variants (cf. I.1.8). For Rosser's original aim in [20], the elimination of the assumption of ω-consistency (or rather 1-consistency in the sense of [16]) from Gödel's *first* incompleteness theorem, it was sufficient to find a formula A_R such that

$$\vdash_{\mathscr{F}}[\mathrm{Con}_{\mathscr{F}} \to \neg(\vdash_{\mathscr{F}} A_R)] \quad \text{and} \quad \vdash_{\mathscr{F}}[\mathrm{Con}_{\mathscr{F}} \to \neg\vdash_{\mathscr{F}}(\neg A_R)].$$

(As in I.1.8, for a Gödel sentence A_G of a 'usual' $\mathscr{F}$ we do not have $\vdash_{\mathscr{F}}[\mathrm{Con}_{\mathscr{F}} \to \neg\vdash_{\mathscr{F}}(\neg A_G)]$.) Rosser's solution, in the terminology of the present paper, was to pass from $\mathscr{F}$ to a new system, say $\mathscr{F}_R^{(i)}$, and consider Gödel

sentences of $\mathscr{F}_R^{(i)}$. The formal rules determining $\mathscr{F}_R^{(i)}$ are just those of $\mathscr{F}$ except for an additional requirement:

> A derivation d of $\mathscr{F}$, with the end formula A_d, is accepted in $\mathscr{F}_R^{(i)}$ just in case $A_{d'}$ is distinct from $\neg A_d$ for all d' preceding d.

This addition to $\mathscr{F}$ is less symmetric than ours for $\mathscr{F}_R$ of I.1.8, call it $\mathscr{F}_R^{(ii)}$ for the time being, where A_d is not permitted to be of the form $\neg A_{d'}$ either. More specifically, we do not know whether $\mathscr{F}_R^{(i)}$ (or, equivalently $\mathscr{F}$ for consistent $\mathscr{F}$) *formally proves the consistency of* $\mathscr{F}_R^{(i)}$. We know little about $\mathscr{F}_R^{(i)}$ (and perhaps want to know less): Does it satisfy I.1.2? or I.1.3? (demonstrable completeness for Σ_1^0 sentences, resp. demonstrable closure under *modus ponens*), e.g., if $\mathscr{F}$ is $\mathscr{U}\mathscr{A}$ or $\mathscr{C}\mathscr{F}\mathscr{A}$.

In contrast, we know that $\mathscr{F}_R^{(ii)}$ does not satisfy I.1.2, by [9], since $\vdash_{\mathscr{P}\mathscr{R}\mathscr{A}} \mathrm{Con}_R^{(ii)}$. Also, by I.2.8 of the present paper, we know that $\mathscr{F}_R^{(i)}$ is not demonstrably closed under *modus ponens* when $\mathscr{F}$ is $\mathscr{C}\mathscr{F}\mathscr{A}$ (but we do *not* know this if $\mathscr{F}$ is $\mathscr{U}\mathscr{A}$).

There is a third variant of Rosser's original solution, e.g., in [8], p. 299 (modified from [11], p. 154, 3.221), call it $\mathscr{F}_R^{(iii)}$. Here a derivation d of $\mathscr{F}$ is accepted if and only if $(\forall d' \leqslant d)\,(\forall d'' \leqslant d)\,(A_{d'} \neq \neg A_{d''})$.

$\mathscr{F}_R^{(iii)}$ has a memorable property, apparently not shared by $\mathscr{F}_R^{(i)}$ nor $\mathscr{F}_R^{(ii)}$: (it can be proved in $\mathscr{P}\mathscr{R}\mathscr{A}$ that)

> If $\mathscr{F}_R^{(iii)}$ contains infinitely many derivations at all, then $\mathscr{F}$ itself is consistent.

This is evident since, if

$$(\exists d' < d)\,(A_{d'} = \neg A_d \vee A_d = \neg A_{d'}),$$

then *all* derivations of $\mathscr{F}_R^{(iii)}$ precede d. Consequently, $\mathscr{F}_R^{(iii)}$ is not demonstrably closed under substitution (not even substitution of equals by equals) nor uniformly complete for *any* formula P in which the relevant parameter n actually occurs and for which each numerical substitution instance is derivable in $\mathscr{F}$; the formulae $P[n/0]$, $P[n/s0]$, etc. would necessarily have (infinitely many) distinct derivations. (We do not know whether $\mathscr{F}_R^{(iii)}$ is demonstrably, locally closed under *modus ponens* when $\mathscr{F}$ is $\mathscr{C}\mathscr{F}\mathscr{A}$; in contrast to the case of $\mathscr{F}_R^{(ii)}$ by the Corollary in I.2.8.2).

Discussion. We recognize the temptation — to some — of looking for 'deeper' significance in Rosser variants because they are obviously related to the so-called empiricist 'philosophy' of mathematics à la Wittgenstein, which questions our understanding of the notion of *application* of a (formal) rule: When — what we normally understand as — correct applications lead to contradictions, we are invited not to reject the rule, but to reject our idea of correct application of the rule; in short, we don't apply the rule and thus avoid inconsistencies. The Rosser variants provide

a neat model of Wittgenstein's speculations. (Though one of us knew these speculations quite well, he was not *consciously* helped by them to construct the Rosser variants which establish their own consistency. As so often, we find here that philosophy *ought* to have been heuristically useful, but we discover this only afterwards.)

We too expect Rosser variants to be useful, but the use we envisage is, admittedly, less direct (and therefore requires more imagination). We believe there must be a genuine theory somewhere between the general theory of arbitrary formal rules, that is (classical), recursion theory on the one hand, and the study of quite special systems such as first order predicate calculus with or without cut, or $\mathcal{U}\mathcal{A}$ and $\mathcal{C}\mathcal{F}\mathcal{A}$, in short, the collection of isolated methods which make up present day proof 'theory' on the other. We surely have enough experience of proofs to judge pretty well which formal rules are definitely relevant or definitely irrelevant for a genuine proof *theory*. But for progress it seems necessary to refine this judgment. We expect that — as so often in similar situations — the study of such border line cases as Rosser variants may well be essential not as an end in themselves but as a tool.

Finally, to turn from expectations to existing results: it is hardly conceivable that it should be worth asking of *some* metamathematical property of a system $\mathcal{F}$, namely $\mathrm{Con}_{\mathcal{F}}$, whether it is formally derivable in $\mathcal{F}$, but not of any others! (such as closure under *modus ponens*). Rosser variants $\mathcal{F}_R$ certainly help to free us of the fixation on consistency, since, as we have seen, Con_R is not problematic.

References

[1] D. de Jongh, *Formulas of one propositional variable in intuitionistic arithmetic*, Fund. Math. (to appear).

[2] S. Feferman, *Arithmetization of metamathematics in a general setting*, ibidem 49 (1960), p. 35–92.

[3] J-Y. Girard, *Une extension de l'interprétation de Gödel à l'analyse, et son application à l'élimination des coupures dans l'analyse et la théorie des types*, Proc. second Scand. Logic Symp., Amsterdam 1971, p. 63–92.

[4] — *Three-valued logic and cut elimination: the actual meaning of Takeuti's conjecture*, to appear.

[5] K. Gödel, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme*. I, Monatsh. Math. Phys. 38 (1931), p. 173–198.

[6] L. Henkin, *Problem*, JSL 17 (1952), p. 160.

[7] D. Hilbert and P. Bernays, *Grundlagen der Mathematik I*, second edition, Berlin 1968.

[8] — *Grundlagen der Mathematik* II, second edition, Berlin 1970.

[9] R. J. Jeroslow, *Redundancies in the Hilbert–Bernays derivability conditions for Gödel's second incompleteness theorem*, JSL 38 (1973), p. 359

[10] G. Kreisel, *On a problem of Henkin's*, Indag. Math. 15 (1953), p. 405–406.

[11] — *Mathematical logic*, in: *Lectures on Modern Mathematics* (ed. T. L. Saaty), vol. III, New York 1965, p. 95–195.

[12] — *A survey of proof theory*, JSL 33 (1968), p. 321–388. ([9] excludes alternative (i) on p. 332 of footnote (6) by showing that Z' does not prove its own consistency.)

[13] — *Church's thesis: A kind of reducibility axiom for constructive mathematics*, Intuitionism and Proof Theory (ed. J. R. Myhill et. al), Amsterdam 1970, p. 121–150.

[14] — *A survey of proof theory* II, Proc. Second Scand. Logic Symp. Amsterdam 1971, p. 109–170.

[15] G. Kreisel and J-L. Krivine, *Elements of mathematical logic*, second revised printing, Amsterdam 1971.

[16] G. Kreisel and A. Levy, *Reflection principles and their use for establishing the complexity of axiomatic systems*, Zeitschr. f. math. Logik und Grundlagen d. Math. 14 (1968), p. 97–142. (Theorem 20 on p. 135–136 can be replaced by the stronger and more elegant result: For systems S considered and k-formulae A: $[(k - \mathrm{Con}S) \wedge S \vdash A)] \to A$, if $k \leqslant 2$. But there is a Σ_3^0 sentence A_0 and an ω-consistent S such that $S \vdash A_0$ and A_0 is false (9).A very much sharpened form of the assertion in the last paragraph of p. 140 has been established in [1]).

[17] G. Kreisel and A. S. Troelstra, *Formal systems for some branches of intuitionistic analysis*, Ann. of Math. Logic 1 (1970), p. 229–387. An addendum (by A. S. Troelstra) ibidem 3 (1971), p. 437–439.

[18] M. H. Löb, *Solution of a problem of Leon Henkin*, JSL 20 (1955), p. 115–118.

(9) This shows that ω-consistency is of little significance for formulae A not in Π_3^0, and 2-cosistency is sufficient to ensure the truth of (derivable) A which are in Π_3^0.

[19] A. Mostowski, *Sentences undecidable in formalized arithmetic*, Amsterdam 1954.
[20] J. B. Rosser, *Extensions of some theorems of Gödel and Church*, JSL 1 (1936), p. 89–91.
[21] W. W. Tait, *A non-constructive proof of Gentzen's Hauptsatz for second order predicate logic*, Bull. Amer. Math. Soc. 72 (1966), p. 980–983.
[22] G. Takeuti, *On a generalized logical calculus*, Japan. J. of Math. 23 (1953), p. 39–96.
[23] — *Consistency proofs of subsystems of classical analysis*, Ann. of Math. (2), 86 (1967), p. 299–348.