Interpretability in Set Theories

Robert M. Solovay

A letter to Petr Hájek, Aug. 17, 1976

Annotation

This is a scan, created in October 2007 and updated in June 2022, of a letter in Petr Hájek's personal archive. The letter was written as a reaction on a question raised in [HH72] whether there exists a set sentence φ such that (GB, φ) is interpretable in GB, but (ZF, φ) is not interpretable in ZF. The proof contained in this letter was never published.

[Sol76b] is a self-citation to *this* file. [Sol76a] is another letter sent earlier in the same year. It solves other problem listed in [HH72], and it was also never published.

R. M. Solovay's postscript note, Oct. 10, 2007

It seems to me that the formulation of the notion of "satisfactory" in section 3 is not quite right. I would rewrite part 3 as follows:

If φ is one of the following sorts of sentence then $s(\varphi) = 1$:

(a) The closure of one of the axioms of $\mathrm{ZF} + \mathrm{V{=}L}$;

(b) The closure of a logical or equality axiom;

(c) One of the special axioms about the c_j's.

--Bob Solovay

References

[HH72] M. Hájková and P. Hájek. On interpretability in theories containing arithmetic. *Fundamenta Mathematicae*, 76:131–137, 1972.

[Sol76a] R. M. Solovay. On interpretability in Peano arithmetic. Unpublished letter to P. Hájek, www.cs.cas.cz/~hajek/RSolovayIntpPA.pdf, May 31, 1976.

[Sol76b] R. M. Solovay. Interpretability in set theories. Unpublished letter to P. Hájek, www.cs.cas.cz/~hajek/RSolovayZFGB.pdf, Aug. 17, 1976.

File created by Zuzana Haniková, Dagmar Harmancová and Vítězslav Švejdar in June 2022. A BibTEX entry to cite this letter can be as follows:

```
@Preamble(
\providecommand{\href}[2]{#2}
\providecommand{\nolinkurl}[1]{\url{#1}}
\newcommand{\rurl}[1]{\href{http://#1}{\nolinkurl{#1}}}
)
@Unpublished(solo:inte76,
  author="Robert M. Solovay",
  title ="On Interpretability in {P}eano Arithmetic",
  note  ="Unpublished letter to P. H{\'a}jek,
          \rurl{www.cs.cas.cz/~hajek/RSolovayZFGB.pdf}",
  year  ="Aug.\@~17, 1976"
)
```

Aug. 17, 1976

Dear Professor Hajek,

I can now settle another question raised in your paper on interpretations of theories. There is a Π^0_1 sentence, Φ, such that

1) $ZF + \Phi$ is not interpretable in ZF.

2) $GB + \Phi$ is interpretable in GB.

Φ will be a variant of the Rosser sentence for GB. However, for my proof to work, I need a "non-standard formalization of predicate logic" (roughly that given by Herbrand's theorem.) I also have to be a bit more ~~picky~~ careful about the Gödel numbering used than is usually

~~necess~~ necessary.

1. Let me begin with the formal language $\mathcal{L}$. Well-formed formulas of $\mathcal{L}$ will consist of certain of the strings on the finite alphabet Σ:

$$\Sigma = \{\ \&,\ \neg,\ \forall,\ \vee,\ (,\ ),\ c,\ \varepsilon,\ =,\ 1,\ 0\ \}$$

To each string on Σ we correlate a number base 12 in ~~decimal~~ notation via $\& \sim 1$, $\forall \sim 3$, etc.

This number is the Gödel number of the symbol

We have in our language an infinite stock of variables $v_0, v_1, v_2, \ldots$, and an infinite string of constants $c_0, c_1, c_2, \ldots$.

For example c_5 will be the string

$$c\,(1\,0\,1).$$

2. I next wish to introduce a theory, T, in the language $\mathcal{L}$. Basically, T is the theory ZFC + V = L. However, to each ~~formula~~ a sentence φ of the form

$$(\exists x)\, \Psi(x)$$

with Gödel number, e, we assign the following axioms:

1) $(\exists x)\, \Psi(x) \rightarrow \Psi(c_e)$

2) $\neg (\exists x)\, \Psi(x) . \rightarrow c_e = 0$

3) $(\forall y)\, [\, y <_L c_e \rightarrow \neg \Psi(y)\,]$

4) $c_e = 0$. (if e is not a Gödel no of the stated form.)

Thus c_e is the least x such that $\Psi(x)$ in the canonical well-ordering of L, ~~other~~ if such an x exists; otherwise $c_e = 0$.

Note that φ may well contain some c_j's, though since $\# \varphi = e$, c_e does not appear in φ.

Our Gödel numbering has been arranged so that:

Let $\varphi(x)$ be a formula. Suppose

$$\log_{12} \# \varphi(x) \leq z,$$

$$\log e \leq z.$$

(Here $\# \varphi$ is the Gödel number of φ.)

Then $\log \# \varphi(c_e) \leq P(z)$, for some explicit polynomial P.

$P(z) = z(z+6)$

3. Let s be a sequence of zero's and one's. $s : m \to 2$, say. s is <u>satisfactory</u> if

1) $s(\# \neg \varphi) = \neg\, s(\# \varphi)$

2) $s(\# \varphi \,\&\, \psi) = s(\# \varphi) \,\&\, s(\# \psi)$

3) If φ is an axiom of $ZFC + V = L$ or

one of the special axioms about the c_j's, then

$$s(\#\varphi) = 1.$$

Of course these conditions only apply for places where s is defined.

We say a sentence Θ is proved at level n if
1) $n > \#\Theta$ and
2) every $s : n \to 2$ which is satisfactory has $s(\#\Theta) = 1$.

It is not hard to show the following are equivalent (for Θ a sentence containing no c_j's):

1) $ZFC + V = L \vdash \Theta$

2) For some n, Θ is proved at level n.

1) ⇒ 2) [illegible] $T_{c_j} \vdash \Theta$ – [illegible] $(ZFC+V=L)_c \vdash \Theta$
2) ⇒ 1) – [illegible] $(ZFC+V=L)_c \vdash \Theta$.

Also note that the relation: "Θ is proved at level n" is primitive recursive, and in fact is

Kalmar elementary.

4. We can now define our variant of the Rosser sentence, Φ: Φ says "If I am proved at level n, then my negation is proved at some level $j \leq n$."

Φ has the usual properties of the Rosser sentence. In particular:

1) Φ is Π^0_1.

2) Φ is undecidable in $ZFC + V = L$.

3) $\vdash Con(GB) \rightarrow \Phi$. (The proof can be carried out in Peano arithmetic.)

It follows from 1) and 2) that ~~Φ is~~ $ZF + \Phi$ not interpretable in ZF. We shall show that

$GB + \Phi$ is interpretable in GB. For that it suffices to show $GB + \Phi$ is interpretable in $GB + \neg\Phi + V{=}L$. We work from now on in the theory $GB + \neg\Phi + V{=}L$.

5. Since $\neg\Phi$ is true, Φ must have been proved at some level n. Let n_0 be the least level at which Φ is proved. (Note that for any standard integer k, $n_0 > k$, though this can only be formulated as a scheme.)

6. An important role in our proof is played by the notion of partial ~~truth~~ satisfaction relation. We begin with some preliminary definitions.

Let j be an integer. If j is the Gödel number of a ~~&~~ well-formed formula, φ, then

A_j is the set of free variables ~~and constants~~ of φ. Otherwise

$A_j = \emptyset$. Let D_j be the class of all ordered

pairs $\langle k, u \rangle$ such that

1) $k < j$

2) k is the Gödel number of a well-formed formula.

3) u is a set.

4) u is a function with domain $A_{\cancel{j}}^{k}$

The following can easily be formalized in

GB: Z is a ~~proper~~ class and is a function

and mapping all constants c_i $B c_i < j$ into V.

mapping D_j into $\{0, 1\}$. We interpret $Z(\langle k, u \rangle) = \varepsilon$

as meaning: if the free variables of φ are interpreted according to u, then $\varphi(u)$ has truth value ε. (Here $\#\varphi = k$.) Finally Z satisfies the usual Tarski inductive definition of truth in so far as they make sense (i.e., in so far as $Z(\langle k, u\rangle)$ is defined.) (in the structure $\langle V, \varepsilon\rangle$, V the class of all sets.)

Let $Tr(j, Z)$ be the formula of GB expressing all this. Then the following are easy to establish:

1) $\vdash_{Z}$ $(\forall j)(\forall Z)(\forall Z')\; Tr(j, Z) \;\&\; Tr(j, Z') \rightarrow Z = Z'.)$

2) $(\forall j)(\forall Z)(\forall k)\;[Tr(j, Z) \;\&\; k < j. \rightarrow (\exists Z')\; Tr(k, Z')$.

3) $(\forall j)(\forall Z)[Tr(j, Z) \rightarrow (\exists Z')\; Tr(j+1, Z')]$.

7. Let $I_0 = \{j : (\exists Z)\, Tr(j, Z)\}$. Our next goal is to show $2^{n_0} \notin I_0$. The reason for 2^{n_0} rather than n_0 is that we intend to use the following lemma.

Let φ be a ~~formula~~ sentence of $\mathcal{L}$ containing the constants $c_{i_1}, \dots, c_{i_k}$. Let $v_1', \dots, v_k'$ be the first k distinct variables not appearing in φ. Let φ' be the formula obtained by replacing c_{i_k} by v_k' in φ. Then if $\# \varphi < n_0$, $\# \varphi' < 2^{n_0}$.

(2^{n_0} could be replaced by $n_0^{\log \log n_0}$, if we desired.)

Let then $Tr(2^{n_0}, Z)$. Using Z we can compute the correct value of c_i (call it $\tilde{c}_i$) for $i < n_0$.

We can then determine the map $s: n_0 \to 2$ that ~~is~~ represents the "true" state of affairs (true according to $\mathfrak{Z}$), interpreting c_i as $\tilde{c}_i$.) This s will be satisfactory and since Φ is false (we are working in $\mathfrak{Z}GB + \neg\Phi + V=L$!), $s(\#\Phi) = 0$. But this contradicts Φ being proved at level n_0.

8. Our next goal is to define a ~~set~~ collection I of integers with the following properties:

1) $4 \in I$

2) Let $z \in I$. Let

$$\log_2 x \le (\log_2 z)^2$$

Then $x \in I$. 3) $n_0 \notin I$.

(I is, like I_0, a definable collection of integers but not a set.) It follows from 1), 2) that I contains all the standard integers and is closed under $+$, $\cdot$, is an initial segment of the integers. Finally, $x \in I$ implies $x^{\log_2 x} \in I$.)

Let $I_1 = \{m : (\forall n \in I_0)(m + n \in I_0)\}$.

Then $I_1 \subseteq I$, and I_1 is an initial segment of the integers closed under $+$.

Let $I_2 = \{m : 2^m \in I_1\}$.

Then I_2 is closed under $+1$, is an initial segment of I_0 and does not contain n_0.

Repeat the process by which I_2 was obtained from I_0 three times ~~twice~~, getting I_8 such that I_8 is an init seg of

ω, closed under $+1$, and such that

$$x \in I_8 \to 2^{2^{2^x}} \in I_2.$$

Let $I = \{z : (\exists x \in I_8)\ z \leq 2^{2^{2^x}}\}$. Then I has the stated properties.

Now since $n_0 \notin I$, $n_0 - 1 \notin I$. Let s be the least satisfactory map of $n_0 - 1$ into 2 such that $s(\#\Phi) = 1$. (s exists, since otherwise $\neg\Phi$ would be proved at level $n_0 - 1$, and Φ would be true. (We are using that $\#\neg\Phi < \cancel{\#} n_0$ since $\#\neg\Phi$ is standard.)) We are going to use s to define an interpretation of $GB + \Phi$.

It will be tacitly assumed that all the sentences we form have Gödel numbers in I. This may be proved using the closure properties of I.

We first define an equivalence relation $\sim$ on I. $i \sim j$ if $s(c_i = c_j) = 1$. Each $\sim$-class has a least member (since s is a set!). Let

$$M = \{ x \in I : (\forall y \in I)(y \sim x \rightarrow x \leq y) \}.$$

We put an ε-relation on M by putting

$$x \,\varepsilon_M\, y \quad \text{iff} \quad s(c_x \,\varepsilon\, c_y) = 1.$$

Then for φ of standard length $s(\varphi(c_{i_1}, \ldots, c_{i_m})) = 1$ iff $\langle M ; \varepsilon_M \rangle \models \varphi(c_{i_1}, \ldots, c_{i_m})$. In particular $\langle M, \varepsilon_M \rangle \models ZF + V = L + \Phi$.

We make M into a model of $\mathfrak{Z}GB$ as follows. Let $S = \{ e \in I :$ e is the Godel no. of a formula having only v_0 free. We define an equivalence relation $\sim_1$ on S by putting $e_1 \sim_1 e_2$ if

$$s(\ (\forall v_0)\ [\varphi_{e_0}(v_0) \Leftrightarrow \varphi_{e_1}(v_0)]\ ) = 1.$$

As before each $\sim_1$ equivalence class has a least element. Let S^* be the set of these $\sim_1$-minimal elements. Define the membership relation between S^* and M via ~~$\in$~~ if

$$j \in e \quad \text{iff} \quad s(\ \varphi_e(c_j)\ ) = 1.$$

Of course $S^* \cap M$ need not be empty. This is handled by replacing S^* by $\{1\} \times S^*$, M by $\{0\} \times M$. We now have a model of $GB + \Phi$ except each set has a copy among the classes. But this minor defect is handled in a well-known way. The upshot is we have interpreted

$$GB + \Phi \quad \text{in} \quad GB + \neg\Phi + V = L.$$

I hope (presuming this is new work) to write up a paper containing this result as well as the one in my earlier letter. When I do, I shall, of course, send you a preprint.

Sincerely yours,

Bob Solovay