ANALYSIS 60.1 JANUARY 2000

A proof of the (strengthened) Liar formula in a semantical extension of Peano Arithmetic

JEFFREY KETLAND

Let *PA* be standard first-order Peano Arithmetic in *L*, the first-order language of arithmetic. Let *PA*(*S*) be a semantical extension of *PA* obtained by adding a primitive satisfaction predicate $Sat_L(x, y)$, governed by Tarskian axioms. Let L_S be the resulting language – i.e., *L* plus the new primitive $Sat_L(x, y)$. The system of axioms governing $Sat_L(x, y)$ was given in Tarski 1936. Anyone interested in delving into some highly technical work on this topic could consult Feferman 1991, Kaye 1991 and Halbach 1999.[1]

The *philosophical* significance of such semantical extensions is discussed in Ketland 1999, where such extensions are compared with *minimalistic* extensions generated by adding just the 'T-scheme' (the set of formulas $Tr_L(\ulcorner \phi \urcorner) \leftrightarrow \phi$, where ϕ does not contain $Tr_L(x)$). These latter extensions are *conservative*. Similar constructions are considered in Shapiro 1998. And, as both Shapiro and I argued, the issue of (non-)conservativeness of adding a theory of truth to a base theory can be related to the issue of *deflationism* about truth.[2]

[1] The notation *PA*(*S*) is due to Kaye 1991. It is worth mentioning that an ambiguity arises concerning what is meant, in defining the satisfaction-theoretic extension, by 'adding' the new axioms which contain *new vocabulary*. The ambiguity is whether to expand the induction scheme in *PA* to include formulas containing $Sat_L(x, y)$. It turns out that if induction is *not* expanded (which is arguably 'unnatural'), then *PA* + satisfaction axioms is a *conservative extension*. However, if induction is expanded, so inductive proofs involving the formulas containing Sat_L and Tr_L can be formalized, then the extension *PA* + satisfaction axioms is a *non-conservative extension*, and indeed, this is the theory we refer to as *PA*(*S*). The non-conservativeness is witnessed by the fact that $PA(S) \vdash Con(PA)$.

(In fact, the system *PA*(*S*) is fully intertranslatable with a certain subsystem of second-order arithmetic called *ACA* ('Arithmetic Comprehension Axiom'). See Halbach 1999).

[2] The crux of the matter is that it can be shown that $PA(S) \vdash True(PA)$ where $True(PA)$ is the 'soundness' formula $\forall x(Prov(x) \rightarrow Tr(x))$ (or 'Global Reflection Principle') and where $Prov(x)$ is the standard provability predicate for *PA*. So, *PA*(*S*) proves that 'any theorem of *PA* is true'. Indeed, an analogous construction can be given for any sufficiently rich formalized system *F* (such as *ZFC*) resulting in a semantical extension

ANALYSIS 60.1 January 2000, pp. 1–4.

Introduce within this theory the truth predicate $Tr_L(x)$, governed by the explicit definition,

(1) $\quad x(Tr_L(x) \leftrightarrow (Sen_L(x) \quad y(Seq(y) \rightarrow Sat_L(x,y))))$

The idea is that $Sat_L(x,y)$ expresses the *satisfaction relation* between (codes of) L-formulas and (codes of) sequences.[3] Then $Tr_L(x)$ expresses the *concept of truth* for such L sentences. It can be shown that this theory satisfies Tarski's Convention T: i.e.,

(2) $\quad PA(S) \vdash Tr_L(\ulcorner \phi \urcorner) \leftrightarrow \phi$

for each closed formula ϕ of L. This syntactic looking fact has a nice model-theoretic corollary. Let (M, S, Tr) be *any* model of $PA(S)$. Let $\#\phi$ be the gödel number of the L-formula ϕ. Then the fact (2) entails that, for any closed L-formula ϕ,

(3) $\quad \#\phi \quad Tr \Leftrightarrow M \vDash \phi$

Furthermore, it can easily be shown that any model of $PA(S)$ will assign to $Tr_L(x)$ *exactly* the (codes of) L-sentences that hold in the reduct M. That is, it gets the *extension* of $Tr_L(x)$ exactly right. It is important to notice the presence of the formula $Sen_L(x)$ in the definition of $Tr_L(x)$. It is this which ensures that *only* (the codes of) L-sentences enter extension of $Tr_L(x)$. Hence,

(4) $\quad$ If $(M, S, Tr) \vDash PA(S)$ then $Tr = \{\#\phi : \boldsymbol{M} \vDash \phi$ and $\phi \quad Sen(L)\}$

To return to the main point, there is an L-formula $Sen_L(x)$ which *strongly defines* within $PA(S)$ the (recursive) set of codes of closed L-formulas. That is,

(5) $\quad$ If ϕ is a closed L-formula, then $PA(S) \vdash Sen_L(\ulcorner \phi \urcorner)$
(6) $\quad$ If ϕ is not a closed L-formula then $PA(S) \vdash \neg Sen_L(\ulcorner \phi \urcorner)$

Furthermore, the definition (1) of $Tr_L(x)$ guarantees that the following holds:

(7) $\quad PA(S) \vdash \quad x(Tr_L(x) \rightarrow Sen_L(x))$

$F(S)$. The analogous result is that $F(S) \vdash$ dash $True(F)$. It is clear that this is a proper extension of F, because $F(S) \vdash Con(F)$. I do not know whether the strength of such systems have been examined in any more detail.

Incidentally, it *is* known that $PA(S) \vdash PH$, the Paris-Harrington formula which is famously not a theorem of PA. See Kaye 1991.

[3] Normally, the metatheory requires a certain amount of set theory. However, only finite sequences are needed to recursively define satisfaction for first-order languages, all of whose formulas contain only finitely many variables. The reason is that the class of finitesequencesofnaturalnumbersiscountableandrecursive:thuseachsuchsequence can be coded (in an effective manner) as a number. Then a predicate $Seq(x)$ strongly defining this set can be defined in PA. Thus, we assume that the sole new concept introduced into $PA(S)$ is the concept expressed by satisfaction predicate $Sat_L(x, y)$.

Next, think about the Diagonalization Lemma (or Fixed Point Theorem). The system $PA(S)$ satisfies the requirements of this theorem (it is an extension of Robinson Arithmetic Q). Thus, there must be a fixed point formula λ such that,

(8) $PA(S) \vdash \lambda \leftrightarrow \neg Tr_L(\ulcorner \lambda \urcorner)$

The analysis of the proof of the Diagonalization Lemma[4] shows that this formula λ must contain the new predicate $Tr_L(x)$. Indeed, because the truth-in-L predicate $Tr_L(x)$ cannot, by Tarski's Indefinability Theorem, be defined in PA, it follows that such a formula is not a closed L-formula (and is not logically equivalent to or logically interdeducible with any L-formula). Thus, λ must be an L_S formula.

This formula λ is the formal analogue of the so-called 'strengthened liar' for our formalized semantical system $PA(S)$. It is a formula that 'says of itself that it is not true'. But what does 'true' mean here? Well, it has to mean 'true-in-L'. Intuitively, this claim is, in fact, correct: i.e., λ is, in fact, *not* true-in-L. Thus, λ is, in fact, *true* (in L_S). This is, in effect, Tarski's own resolution of the liar paradox (including the strengthened liar).

We shall use the above facts (5) – (8) to deduce something rather interesting. Namely, that the strengthened liar formula λ is a *theorem* of $PA(S)$. Thus, λ is provably true-in-L_S! The required proof that the formula λ is a theorem of $PA(S)$ is triviality itself. Since, as we noted, λ is not a closed L-formula, we can deduce from (6) that,

(9) $\vdash PA(S) \vdash \neg Sen_L(\ulcorner \lambda \urcorner)$

It immediately follows from (7) that,

(10) $PA(S) \vdash \neg Tr_L(\ulcorner \lambda \urcorner)$

and thus, from (8),

(11) $PA \vdash \lambda$

This interesting result deserves further comment. Intuitively, $PA(S)$ is in fact *true* (it is certainly true in the standard expansion $(\mathfrak{N}, S, Tr)$[5] of the

[4] See, e.g., (Boolos Jeffrey 1989). If ϕ is any formula, then the *diagonalization* of ϕ, call it Diag(ϕ), is the formula $\exists x(x = \ulcorner \phi \urcorner \wedge \phi)$. (Note that if ϕ contains the variable x free, then Diag(ϕ) is equivalent to the formula $\phi(\ulcorner \phi \urcorner)$ which obviously says that ϕ *is true of the code of itself*). Taking codes, we get the diagonal function *diag* on numbers. This function is provably recursive. So, suppose that the L-formula $Diag(x, y)$ represents (in our theory $PA(S)$) this diagonal function *diag*. Next let Φ be the L_S-formula $\exists y(Diag(x, y) \wedge \neg Tr_L(y))$. Finally, let λ be Diag(Φ). It then quickly follows that $\lambda \leftrightarrow \neg Tr_L(\ulcorner \lambda \urcorner)$ is a theorem of $PA(S)$.

[5] S is the *satisfaction relation* on the standard model $\mathfrak{N}$. I.e., it is the set of pairs (n, m) where n is the code of an L-formula ϕ and m is the code of some finite sequence s and $\mathfrak{N} \models_s \phi$; and Tr is the corresponding set of codes of true formulas (in $\mathfrak{N}$).

intended structure $\mathfrak{N}$). It follows that λ is, in fact, true (or, if you like, it holds in $(\mathfrak{N},S,Tr)$). Furthermore, it follows that $\neg Tr_L(\ulcorner\lambda\urcorner)$ is also true! This correctly expresses the fact that the L_S-formula λ is in fact not true in L: it is not true in L for the rather trivial 'syntactic' reason that λ is not equivalent to any L-formula. Indeed, it turns out that the λ's code – the number #λ – cannot be an element of Tr, for this set is constrained by the *definition* of $Tr_L(x)$ given above ((1)) to be a *subset* of the set (of codes) of closed L-formulas. The formula λ is the formal equivalent for our formal system $PA(S)$ of the strengthened liar, which allegedly 'says of itself that it is not a true sentence'. But, in our study, this claim is actually true. The strengthened liar sentence λ is not even a sentence of L, and is *a fortiori* not a *true* sentence of L. However, λ is, in fact, a *true* sentence of the extended language L_S and is, in fact, *provable* in $PA(S)$.

The London School of Economics
Houghton Street, LONDON WC2A 2AE
j.j.ketland@lse.ac.uk

References

Boolos, G. and R. Jeffrey. 1989. *Computability and Logic*. Third edition. Cambridge: Cambridge University Press.

Feferman, S. 1991. Reflecting on incompleteness. *Journal of Symbolic Logic* 56: 1–49.

Halbach, V. 1999. Conservative theories of classical truth. *Studia Logica* 62: 353–70.

Kaye, R. 1991. *Models of Peano Arithmetic*. (Oxford Logic Guides 15). Oxford: Clarendon Press.

Ketland, J. 1999. Deflationism and Tarski's paradise. *Mind* 108: 69–94.

Shapiro, S. 1998. Truth and proof – through thick and thin. *Journal of Philosophy* 95: 493–522.

Tarski, A. 1936. Der wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* 1: 261–405. English translation, by J. H. Woodger, 'The concept of truth in formalized languages', appeared in A. Tarski 1956, *Logic, Semantics and Metamathematics: Papers by Alfred Tarski from 1923 to 1938*. Oxford: Clarendon Press.