Logic and Logical Philosophy
Volume 20 (2011), 205–214
DOI: 10.12775/LLP.2011.011

Josef Wolfgang Degen

# SOCRATES DID IT BEFORE GÖDEL

**Abstract.** We translate Socrates' famous saying *I know that I know nothing* into the arithmetical sentence *I prove that I prove nothing*. Then it is easy to show that this translated saying is formally undecidable in formal arithmetic, using Gödel's Second Incompleteness Theorem. We investigate some variations of this Socrates-Gödel sentence. In an appendix we sketch a ramified epistemic logic with propositional quantifiers in order to analyze the Socrates-Gödel sentence in a more logical way, separated from the arithmetical context.

**Keywords**: The "paradoxon" of Socrates, Gödel's Second Incompleteness Theorem. A ramified epistemic logic with propositional quantifiers.

## 1. Introduction

Socrates used to say *I know that I know nothing*, thereby implying that he was superior to his contemporaries who did not even know that they knew nothing.

However, is this saying of Socrates not self-contradictory? For, on the one hand, he knows something, namely that he knows nothing. And by the same token, he knows nothing.

The purpose of this short note is to show that Socrates' saying is far from self-contradictory if formalized carefully. Quite the contrary, its careful formalization is the first formally undecidable arithmetical proposition. In this sense, Socrates did "IT" before Gödel; and that Socrates did it before Gödel was, of course, confirmed and even proved by Gödel (in 1931).

That a person $A$ knows a proposition $\varphi$ means that $A$ possesses a theory $\Sigma$ in which $\varphi$ is provable, in symbols: $\Sigma \vdash \varphi$. Thus the reference

*Received* December 17, 2009; *Revised* March 2, 2011
 ISSN: 1425-3305

to $A$'s Ego as a bearer of $A$'s knowledge is immaterial.[1] Then Socrates' saying translates into the assertion: *It is provable in* $\Sigma$ *that nothing is provable in* $\Sigma$. For a wide range of theories $\Sigma$ this assertion is an arithmetical sentence, and it is neither provable nor refutable in $\Sigma$.

## 2. A Formal Execution

In the following exposition we will follow rather closely the presentation in [2], which the reader may look up for more details.

We assume that $\Sigma$ is recursively axiomatized, containing enough[2] arithmetic about the natural numbers 0, 1, 2, 3, ... and every arithmetical sentence $\varphi$ provable in $\Sigma$ is true in the standard natural numbers $\omega = \{0, 1, 2, \ldots\}$ together with addition and multiplication. We denote this standard model by $\langle \omega, 0, 1, +, \times \rangle$ or simply by $\langle \omega \rangle$. Hence $\Sigma$ is consistent. (Truth in the standard natural numbers is a stronger property than mere consistency.)

As usual, we assign in an effective and injective way natural numbers $\sharp\alpha$ to expressions $\alpha$ of the language of $\Sigma$. Such natural numbers $\sharp\alpha$ are called Gödel numbers. Thus, expressions are coded by natural numbers, which is by now an ubiquitous and familiar procedure also in theoretical computer science.

Let's now denote by the arithmetical formula $\sigma(n)$ the standard provability predicate for $\Sigma$. More precisely, one constructs first the arithmetical formula $Bew(y, n)$ with the meaning: $y$ is the Gödel number of a proof in $\Sigma$ of the formula (or of the sentence)[3] with the Gödel number $n$. The formula $\sigma(n)$ is then defined by $\exists y Bew(y, n)$.

Then one can show, amongst other things, the two following facts.

$$\Sigma \vdash \varphi \iff \Sigma \vdash \sigma(\sharp\varphi). \tag{1}$$

(For the direction from right to left we use truth in the standard natural numbers. For the direction from left to right we use the numeralwise representability of every primitive recursive relation in $\Sigma$.)

$$\Sigma \vdash \ \sigma(\sharp(\varphi \to \psi)) \to (\sigma(\sharp\varphi) \to \sigma(\sharp\psi)). \tag{2}$$

[1] Similarly, if the sentence $\varphi$ stands in a book, then it is immaterial for the truth or falsity of $\varphi$ whether or not $A$ is the owner of the book.

[2] E.g. primitive recursive arithmetic is more than enough.

[3] A sentence is a formula without free variables.

After these preliminaries, it is clear that Socrates' saying *I know that I know nothing* translates (in accordance to our informal considerations above) into

$$[socrates] \quad :\Longleftrightarrow \quad \sigma(\sharp\forall x \neg\sigma(x)).^{4} \tag{3}$$

THEOREM 1. (a) $\Sigma \nvdash [socrates]$ (b) $\Sigma \nvdash \neg[socrates]$

That is to say, the arithmetical sentence $[socrates]$ is formally undecidable in $\Sigma$, or to express it in a different way: the sentence $[socrates]$ is a witness to the incompleteness of the theory $\Sigma$.

PROOF. *Ad* (a) Suppose $\Sigma \vdash [socrates]$. Then by the fact (1) and definition (3) we get $\Sigma \vdash \forall x \neg\sigma(x)$. Substituting for $x$ the Gödel number $\sharp\forall x\neg\sigma(x)$ we get $\Sigma \vdash \neg\sigma(\sharp\forall x\neg\sigma(x))$. That is the same as $\Sigma \vdash \neg[socrates]$. A contradiction to the consistency of $\Sigma$.

*Ad* (b) Now suppose $\Sigma \vdash \neg[socrates]$. We shall deduce from this that $\Sigma \vdash \neg\sigma(\sharp(0=1))$, that is $\Sigma$ proves its own consistency; and that is impossible by Gödel's Second Incompleteness Theorem (which Socrates did not know, of course, since he knew nothing!)

First we have $\Sigma \vdash 0 = 1 \rightarrow \forall x\neg\sigma(x)$ by the logical principle *ex falso quodlibet*. By (1) we have $\Sigma \vdash \sigma(\sharp(0 = 1 \rightarrow \forall x\neg\sigma(x)))$. By another easily verifiable law concerning the provability predicate, namely fact (2) above, the last fact yields

$$\Sigma \vdash \sigma(\sharp(0=1)) \rightarrow \sigma(\sharp\forall x\neg\sigma(x)).$$

By contraposition and definition we have

$$\Sigma \vdash \neg[socrates] \rightarrow \neg\sigma(\sharp(0=1)).$$

Hence by applying the *modus ponens* we get the desired

$$\Sigma \vdash \neg\sigma(\sharp(0=1)). \qquad \dashv$$

So we may ask which of the sentences $[socrates]$, $\neg[socrates]$ is true in the standard natural numbers?

The answer is left to the reader. If the reader can give no immediate intuitive argument (s)he should first prove $\Sigma \vdash [socrates] \leftrightarrow \sigma(\sharp(0=1))$. Not totally easy. Try your hands on it before reading the next passage.

[4] We should correctly write $\forall x(sent(x) \rightarrow \neg\sigma(x))$ instead of $\forall x\neg\sigma(x))$ where $sent(x)$ means that $x$ is the Gödel number of a sentence. However, we shall mostly omit such relativization of quantifiers to avoid cluttering up our notations.

If found, the solution is easy to understand. The direction $\leftarrow$ has already been proved. For $\rightarrow$ start with $\Sigma \vdash 1 = 1$; hence $\Sigma \vdash \sigma(\sharp(1 = 1))$. From this we get $\Sigma \vdash \exists x \sigma(x)$. This in turn is logically equivalent to $\Sigma \vdash \forall x \neg \sigma(x) \rightarrow 0 = 1$. Finally, by (1) and (2) we get $\Sigma \vdash \sigma(\sharp \forall x \neg \sigma(x)) \rightarrow \sigma(\sharp(0 = 1))$.

*Remark.* The phenomenon described in the present paper in an arithmetical context gives rise to a ramified epistemic logic with propositional quantifiers. In this logic Socrates' saying, its negation, and related propositions can be thoroughly discussed. See Appendix.

The next two sections contain not much in the way of theory, but several exercises which the reader might find interesting and inspiring.

## 3. Some Variations

In the sentence $\sigma(\sharp \forall x \neg(\sigma(x))$ there are two occurrences of the provability predicate $\sigma(n)$, the first positive, the second negative. There are four such possibilities, namely

| | | |
|---|---|---|
| $(+\forall+)$ | $\sigma(\sharp \forall x \sigma(x))$: | *I know that I know everything* |
| $(+\forall-)$ | $\sigma(\sharp \forall x \neg \sigma(x))$: | *I know that I know nothing* |
| $(-\forall+)$ | $\neg \sigma(\sharp \forall x \sigma(x))$: | *I do not know that I know everything* |
| $(-\forall-)$ | $\neg \sigma(\sharp \forall x \neg \sigma(x))$: | *I do not know that I know nothing* |

What about provability or nonprovabiliy of these four sentences in $\Sigma$? Sentence $(+\forall-)$ is formally undecidable as already shown; hence its negation $(-\forall-)$ is formally undecidable too.

Consider sentence $(+\forall+)$. It cannot be proved in $\Sigma$. For suppose $\Sigma \vdash (+\forall+)$. Then by (1)[from right to left] we have $\Sigma \vdash \forall x \sigma(x)$. Then $\Sigma \vdash \sigma(\sharp(3 = 1))$. Again by (1)[from right to left] we have $\Sigma \vdash 3 = 1$. And $\Sigma$ is inconsistent since obviously $\Sigma \vdash \neg\, 3 = 1$.

Next we show that also $\Sigma \nvdash (-\forall+)$. Suppose $(*)\ \Sigma \vdash \neg \sigma(\sharp \forall x \sigma(x))$. We have $\Sigma \vdash 0 = 1 \rightarrow \forall x \sigma(x)$. Then by (1) and (2) we get $\Sigma \vdash \sigma(\sharp(0 = 1)) \rightarrow \sigma(\sharp \forall x \sigma(x))$. By contraposition and *modus ponens* together with $(*)$ we get the contradiction $\Sigma \vdash \neg \sigma(\sharp(0 = 1))$.

*Summary.* Each of the four sentences of our above list is formally undecidable.

## 4. Further Variations

Next we make up a new table of four sentences where the quantifier $\forall$ is replaced by the quantifier $\exists$ and see what will happen.

| | | |
|---|---|---|
| $(+\exists+)$ | $\sigma(\sharp\exists x\sigma(x))$: | *I know that I know something* |
| $(+\exists-)$ | $\sigma(\sharp\exists x\neg\sigma(x))$: | *I know that there is something I do not know* |
| $(-\exists+)$ | $\neg\sigma(\sharp\exists x\sigma(x))$: | *I do not know that I know something* |
| $(-\exists-)$ | $\neg\sigma(\sharp\exists x\neg\sigma(x))$: | *I do not know that there is something I do not know* |

*Ad* $(+\exists+)$ First we have $\Sigma \vdash 1=1$. Then by (1) $\Sigma \vdash \sigma(\sharp(1=1))$. Hence by pure logic $\Sigma \vdash \exists x\sigma(x)$. Finally by (1) we get the desired $\Sigma \vdash \sigma(\sharp\exists x\sigma(x))$.

Hence our sentence is provable; therefore

*Ad* $(-\exists+)$ This sentence is refutable (in $\Sigma$).

*Ad* $(+\exists-)$ Suppose $\Sigma \vdash \sigma(\sharp\exists x\neg\sigma(x))$. By (1)[from right to left] we have $\Sigma \vdash \exists x\neg\sigma(x)$. We want to show $(*)$ $\Sigma \vdash \exists x\neg\sigma(x) \to \neg\sigma(\sharp(0=1))$ from which we can deduce by *modus ponens* the contradiction $\Sigma \vdash \neg\sigma(\sharp(0=1))$. However, instead of $(*)$ it is easier to prove its contrapositive

$$\Sigma \vdash \sigma(\sharp(0=1)) \to \forall x\sigma(x)$$

To do this we have to go down to the definition of $\sigma(n)$. We work in $\Sigma$.

So, let $Bew(y,\sharp(0=1))$ for some $y$, and let $a$ be the Gödel number of a sentence, i.e. we suppose $sent(a)$. Then there is a number $z$ such that $sent(a) \to Bew(z,\sharp\alpha)$ where $\alpha$ is the formula $0=1\to\beta$ with $\sharp\beta = a$. It follows that there exists a $v$ such $Bew(v,a)$. Hence we have $sent(a)\wedge Bew(y,\sharp(0=1)) \to Bew(v,a)$. Introducing the quantifiers in the correct order and using the definitions we finally have the desired $\Sigma \vdash \sigma(\sharp(0=1)) \to \forall x\sigma(x)$.

*Ad* $(-\exists-)$ Not provable in $\Sigma$. This can be shown by a meanwhile familiar argument. By the way, although it is seldom explicitly mentioned in the literature about Gödel's Theorems, we have $\Sigma \nvdash \neg\sigma(\sharp\psi)$ for *every* sentence $\psi$, by the meanwhile familiar argument.

*That is, claims of ignorance, i.e. of not-knowing something, are never provable – provided the system in question is consistent.*

## 5. Conclusion

All sentences in the $(*\forall*)$ list are formally undecidable; in the $(*\exists*)$ list just two sentence are formally undecidable; of the remaining two, one is provable and the other is refutable. A remarkable feature of our sentences is that none of them arises by a fixed point construction. Of course, the unprovability of consistency, which is used in our arguments, relies on a fixed point construction. (Recall: the fixed point lemma (or the diagonalization lemma. See [2] for the simple proof.) yields a sentence $\gamma$ such that $\Sigma \vdash \gamma \leftrightarrow \neg\sigma(\sharp\gamma)$. Since we can use the properties (1) and (2) of $\Sigma$ we don't need Rosser's refinement in order to show that $\gamma$ is formally undecidable in $\Sigma$. Using the Bernays derivability conditions one shows $\Sigma \vdash \gamma \leftrightarrow \neg\sigma(\sharp(0 = 1))$. See again [2] for details.

We recommend to the reader to make up and investigate further lists like our two lists $(*\forall*)$ and $(*\exists*)$ about $\sigma(n)$, i.e. of **sentences that do not arise as fixed points**, but whose provability status can be (easily) settled. Try your hand at

$$\forall x(sent(x) \land \sigma(x) \to \exists y(sent(y) \land y > x \land \neg\sigma(y))),$$
$$\sigma(\sharp\exists x(sent(x) \land \forall y(sent(y) \land y > 5x \to \sigma(y)))).$$

Before doing these and similar exercises the reader is invited to read [3].

## Appendix

Here I give a sketch of the promised ramified epistemic logic with propositional quantifiers. This logical system I call SOCR. First two remarks.

(1) The issue of quantifiers over propositions is still in an unsettled state. The intuitionistic sequent calculus with just $\to$ and $\forall$ (binding propositional variables) is very complicated, and the cut elimination theorem for it is a very complicated and strong result.

(2) Ramifications, first used in Principia Mathematica, are by now largely forgotten, and are used only in several systems of predicative second order arithmetic, and there also only as a tool. But in SOCR ramifications are essential and constitutive. There are certainly other purely logical systems in which ramifications are essential and constitutive.

The precise form of SOCR depends on the ramification indices we choose. For definiteness let us use as ramification indices all ordinals strictly smaller than $\omega^\omega$. We call the ordinals below $\omega^\omega$ also *levels*.

**The language of SOCR**

(1) For each level $\alpha$ we have free propositional variables $a^\alpha$, $b^\alpha$, ${a_1}^\alpha$, … and bound (propositional) variables $x^\alpha$, $y^\alpha$, ${x_1}^\alpha$, … of the superscripted level. Optionally we may have for some level some propositional constants of that level.

(2) The logical signs are $\neg$, $\rightarrow$, $\wedge$, $\vee$, $\forall$, $\exists$, $\mathbb{K}$, and the brackets (, ).

The unary connective $\mathbb{K}$ means “He knows that”. If He is Socrates, then in “She knows that” She is perhaps Xantippe.

Now we define by simultaneous induction the formulae and their levels.

(3) Each free propositional variable of level $\alpha$ is a formula of level $\alpha$. Also, each propositional constant of level $\alpha$ is a formula of level $\alpha$.

(4) If $\varphi$ and $\psi$ are formulae of level $\alpha$ and $\beta$, respectively, and if $* \in \{\rightarrow, \wedge, \vee\}$, then $(\varphi * \psi)$ is a formula of level $\max(\alpha, \beta) + 1$. Furthermore, $\neg\,\varphi$ is a formula of level $\alpha + 1$, and $\mathbb{K}\,\varphi$ is a formula of level $\lambda$, where $\lambda$ is the smallest limit ordinal such that $\alpha < \lambda < \omega^\omega$.

(5) If $\mathcal{F}[a^\alpha]$ is a formula of level $\beta$ in which the bound propositional variable $x^\alpha$ does not occur, then $\forall\, x^\alpha\, \mathcal{F}[x^\alpha]$ and $\exists\, x^\alpha\, \mathcal{F}[x^\alpha]$ are formulae of level $\lambda$, where $\lambda$ is the smallest limit ordinal such that $\max(\alpha, \beta) < \lambda < \omega^\omega$.

If $\varphi$ is a formula of level $\gamma$, then we may write this level as superscript, as in the expression $\varphi^\gamma$. A formula without free variables is called a *sentence*.

The calculus SOCR is a sequent calculus. For the connectives $\neg$, $\rightarrow$, $\wedge$, $\vee$ we choose the usual rules. For $\mathbb{K}$ we choose the S4 rules. We have the usual structural rules, and the cut rule of the form

$$\text{(cut)} \quad \frac{\Phi_1 \implies \Psi_1, \varphi \qquad \varphi, \Phi_2 \implies \Psi_2}{\Phi_1, \Phi_2 \implies \Psi_1, \Psi_2}$$

The formula $\varphi$ is called *the cut formula*.

Now it remains to state the rules for the quantifiers. We state all four rules in order to show the mechanics of the ramifications by levels.

$$(\exists \implies) \quad \frac{\mathcal{F}[a^\beta], \Phi \implies \Psi}{\exists x^\alpha \mathcal{F}[x^\alpha], \Phi \implies \Psi} \quad \text{for } \alpha \leq \beta$$

$$(\Longrightarrow \exists) \quad \frac{\Phi \Longrightarrow \Psi, \mathcal{F}[\varphi^\beta]}{\Phi \Longrightarrow \Psi, \exists x^\alpha \mathcal{F}[x^\alpha]} \quad \text{for } \alpha \geq \beta$$

$$(\forall \Longrightarrow) \quad \frac{\mathcal{F}[\varphi^\beta], \Phi \Longrightarrow \Psi}{\forall x^\alpha \mathcal{F}[x^\alpha], \Phi \Longrightarrow \Psi} \quad \text{for } \alpha \geq \beta$$

$$(\Longrightarrow \forall) \quad \frac{\Phi \Longrightarrow \Psi, \mathcal{F}[a^\beta]}{\Phi \Longrightarrow \Psi, \forall x^\alpha \mathcal{F}[x^\alpha]} \quad \text{for } \alpha \leq \beta$$

In the rules ($\exists \Longrightarrow$) and ($\Longrightarrow \forall$) we have to fulfil the so-called *eigenvariable* condition, i.e. the free variable $a^\beta$ must not occur in the conclusion, i.e. under the stroke.

As usual, (formal) proofs are finite rooted trees whose leaves are logical axioms of the form $\varphi \Longrightarrow \varphi$, and the inner nodes (including the root) arise by application of one of the rules. The sequent at the root is called the end sequent; it is *proved* by the proof.

A proof is called *intuitionistic* if in every sequent occurring in it the succedent has at most one formula. The system with intuitionistic proofs may be denoted by IntSOCR, while the system without the restriction to intuitionistic proofs may, for emphasis, be denoted by ClassSOCR (classical SOCR).

THEOREM 2. *The system* ClassSOCR *admits elimination of cuts.*

The proof of this theorem is syntactic where one induction parameter is the level of the cut formula. On the other hand, I could not settle the problem whether the system IntSOCR admits cut elimination. To make cut elimination go through in the intuitionistic case we have perhaps to change the rules for $\mathbb{K}$ a little bit.

For none of the two systems I have developed a semantics.

Now let me close the paper with a discussion of Socrates's saying *I know that I know nothing* in the context of SOCR. The only (logical) axioms of SOCR are sequents of the form $\varphi \Longrightarrow \varphi$. But we can extend SOCR by adding nonlogical axioms. These have the form $\Longrightarrow \varphi$ where $\varphi$ is a sentence.

Let AX be a set of axioms. Then SOCR + AX is inconsistent if we can in SOCR (with cut) derive the empty sequent $\Longrightarrow$ from AX.[5] If SOCR + AX is not inconsistent it is called consistent.

[5] Recall that from $\Longrightarrow$ we can prove *every* sequent by means of the structural rules.

Let $\alpha$ be a level (i.e. an ordinal $< \omega^\omega$). The sentence $socrates(\alpha)$ is the sentence $\mathbb{K}\forall x^\alpha \neg \mathbb{K}\, x^\alpha$, and $\neg socrates(\alpha)$ is of course the sentence $\neg \mathbb{K}\forall x^\alpha \neg \mathbb{K}\, x^\alpha$. Let $AX(\neg socr)$ be the following set of axioms:

$$\{\Longrightarrow \neg socrates(\alpha) \, : \, \alpha < \omega^\omega \}$$

THEOREM 3. *The system SOCR + $AX(\neg socr)$ is consistent.*

Intuitively each

$$\Longrightarrow \neg socrates(\alpha)$$

is a kind of definite consistency assertion, from which we can deduce by $(\Longrightarrow \exists)$ the sequent

$$\Longrightarrow \exists x^\lambda \neg \mathbb{K}\, x^\lambda$$

for a suitable limit ordinal $\lambda$: *There is something I do not know (prove).*

What can we say about sequents of the form $\Longrightarrow socrates(\alpha)$ (without the negation sign $\neg$)?

The reader should convince himself of the consistency of SOCR + $\{\Longrightarrow socrates(0)\}$. But then we have the following catastrophe.

THEOREM 4. *The system SOCR + $\{\Longrightarrow socrates(1)\}$ is inconsistent.*

PROOF. From $\Longrightarrow socrates(1)$ in S4 and by definition we get $\Longrightarrow \forall x^1 \neg \mathbb{K}\, x^1$. From this in SOCR we get $\Longrightarrow \neg \mathbb{K}(a^0 \to a^0)$. Observe that the formula $(a^0 \to a^0)$ gets level 1.

On the other hand, from the logical axiom $a^0 \Longrightarrow a^0$ we get by $(\Longrightarrow \to)$ the sequent $\Longrightarrow (a^0 \to a^0)$. From this by the necessitation rule of S4 (for $\mathbb{K}$) we get $\Longrightarrow \mathbb{K}(a^0 \to a^0)$. ⊣

Of course, we have further inconsistencies if we replace the ordinal 1 by bigger ordinals. All this shows that Socrates' *positive* saying cannot be vindicated in the context of a ramified epistemic logic — while his *negative* saying can be vindicated, even throughout all levels.

Finally, suppose we erase all levels while retaining all rules. We get thereby the system SOCR(unramified). In this system we derive the empty sequent $\Longrightarrow$ from the (unramified) axiom

(0) $\quad \Longrightarrow \mathbb{K}\forall x \neg \mathbb{K}\, x$

as follows.

With an S4 rule we get $\mathbb{K}\,\forall x \neg \mathbb{K}\, x \implies \forall x \neg \mathbb{K}\, x$. With a cut using (0) we get

(1) $\implies \forall x \neg \mathbb{K}\, x$

From the logical axiom

(2) $\neg \mathbb{K}\,\forall x \neg \mathbb{K}\, x \implies \neg \mathbb{K}\,\forall x \neg \mathbb{K}\, x$

we get by an unramified ($\forall \implies$) the sequent

(3) $\forall x \neg \mathbb{K}\, x \implies \neg \mathbb{K}\,\forall x \neg \mathbb{K}\, x$

With a cut from (1) and (3) we get the sequent

(4) $\implies \neg \mathbb{K}\,\forall x \neg \mathbb{K}\, x$

From the sequent (0) we get by a ($\neg \implies$) the sequent

(5) $\neg \mathbb{K}\,\forall x \neg \mathbb{K}\, x \implies$

Finally, a cut from (4) and (5) yields the empty sequent $\implies$.

So, in the unramified epistemic logic, Socrates's saying is really a paradoxon, or more definitely: a self-contradiction.

## References

[1] Kurt Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik* 38 (1931): 173–198.

[2] Craig Smorynski, "The incompleteness theorems", in: Jon Barwise (editor), *Handbook of Mathematical Logic*, North-Holland, 1977.

[3] Craig Smorynski, *Self-Reference and Modal Logic*, Springer, 1985.

Josef Wolfgang Degen
Erlangen University
Department of Computer Science
degen@informatik.uni-erlangen.de