On Modal Systems with Rosser Modalities

Vítězslav Švejdar*

Appeared in M. Bílková and O. Tomala eds., *The Logica Yearbook 2005: Proc. of the Logica 05 Int. Conference*, pp. 203–214, Philosophia Praha, 2006.

1 Introduction

Sufficiently strong axiomatic theories allow for the construction of self-referential sentences, i.e. sentences saying something about themselves. After the Gödel's paper on incompleteness (Gödel, 1931) the *self-reference* method found further applications—some are listed below—and became even more important. Around say 1970 it appeared that the self-reference method was not only a useful tool, but also an interesting field of study: it became clear that reasoning about self-referential sentences could be made more transparent and limitations of the self-reference method could be clarified by using *modal logic*. The connections of meta-mathematics to modal logic, especially after the Solovay's paper (Solovay, 1976), brought traditional modal logic to the attention of more mathematically oriented logicians and constituted a stimulus in modal logical studies.

In this paper we briefly discuss some of the existing modal systems, putting an emphasis on Rosser modalities, also known as *witness comparison* modalities, and present one new system of that kind. First we fix some symbolism and the way we speak about arithmetic and self-reference.

2 Some preliminaries

By *arithmetical language* we mean the language $\{+, \cdot, 0, \mathrm{S}, \leq, <\}$ with two binary function symbols, a constant, a unary function and two binary predicate symbols. Its intended realization, or the *standard model*, is the structure $\mathbf{N} = \langle \mathrm{N}, +^{\mathbf{N}}, \cdot^{\mathbf{N}}, 0^{\mathbf{N}}, \mathrm{s}, \leq^{\mathbf{N}}, <^{\mathbf{N}} \rangle$, where N is the set of all natural numbers (with zero), $+^{\mathbf{N}}$, $\cdot^{\mathbf{N}}$, $\leq^{\mathbf{N}}$, $<^{\mathbf{N}}$ are the usual addition, multiplication, unstrict

*This work is a part of the research plan MSM 0021620839 that is financed by the Ministry of Education of the Czech Republic.

and strict ordering on the set N, $0^{\mathbf{N}}$ is the number zero, and s is the successor function $a \mapsto a+1$. In the sequel we omit the superscript $\mathbf{N}$; so e.g. "+" can be both a symbol and its standard realization, i.e. addition of natural numbers.

An example of an arithmetical formula is $\exists v(v \cdot x = y)$; it can be read the number x is a divisor of the number y and denoted $x \mid y$. Another example, constructed using the formula $x \mid y$, is $\forall z(z \mid x \to z = \mathrm{S}(0) \vee \mathrm{S}(\mathrm{S}(0)) \mid z)$. This formula says that each divisor of x, except the trivial divisor one, is even. Terms like $\mathrm{S}(\mathrm{S}(0))$ are called *numerals;* the numeral $\mathrm{S}(\mathrm{S}(..\mathrm{S}(0)..)$ with n occurrences of the symbol S is denoted $\overline{n}$. In the arithmetical language, numerals make it possible to speak about particular numbers.

We sometimes use the sans-serif font for informal reading of syntactical objects. Whenever this font is used, the reader is expected to think a little bit about the formula (or, sometimes, the proof) it represents.

The sentence $\overline{n} \mid \overline{m}$ resulting by substituting the numerals $\overline{n}$ and $\overline{m}$ for x and y in the formula $x \mid y$ is valid in the structure $\mathbf{N}$ if and only if, in reality, n is a divisor of m. This property of the formula $x \mid y$ is expressed by saying that the formula $x \mid y$ *defines* the divisibility relation in $\mathbf{N}$. One can check that our second example formula $\forall z(z \mid x \to z = \overline{1} \vee \overline{2} \mid z)$ defines the set of all powers of two. So let us choose $\mathsf{Pow}(x)$ as a shorthand for this formula; we have e.g. $\mathbf{N} \models \mathsf{Pow}(\overline{8})$ and $\mathbf{N} \not\models \mathsf{Pow}(\overline{9})$.

We identify formulas and other syntactical objects with their numerical codes under some fixed *coding of syntax.* So if, for example, φ is an arithmetical sentence then $\mathsf{Pow}(\overline{\varphi})$ is a sentence saying that (the numerical code of) the sentence φ is a power of two. We do not (need to) know whether this sentence is valid in $\mathbf{N}$, it may well depend on details of the chosen coding.

If φ is a formula then $\forall v{\leq}z\,\varphi$ is a shorthand for $\forall v(v \leq z \to \varphi)$. Similarly, the formulas $\forall v{<}z\,\varphi$, $\exists v{\leq}z\,\varphi$, and $\exists v{<}z\,\varphi$ have an obvious meaning. The expressions $\forall v{\leq}z$, $\forall v{<}z$, $\exists v{\leq}z$, and $\exists v{<}z$ are *bounded quantifiers.* A formula is *bounded*, or a Δ_0*-formula,* if all its quantifiers are bounded. A formula is a Σ_1*-formula* if it has the form $\exists x\varphi$ where $\varphi \in \Delta_0$; a formula is a Σ*-formula* if it is obtained from bounded formulas using (any number of) conjunctions, disjunctions, bounded quantifiers and existential quantifiers. Since, for example, the divisibility relation is definable also by the formula $\exists v{\leq}y\,(v \cdot x = y)$, we may think that the formula $x \mid y$ is bounded. Similarly, the quantifier $\forall z$ in the formula $\mathsf{Pow}(x)$ can be written as $\forall z{\leq}x$; so $\mathsf{Pow}(x)$ is another example of a bounded formula.

It is not straightforward to *characterize* the sets definable by Δ_0-formulas. However, a characterization of sets definable by Σ-formulas is known. A deep theorem says that, although the set of all Σ_1-formulas is a proper subset of the set of all Σ-formulas, the classes of all sets definable by Σ_1-formulas and by Σ-formulas coincide: a set $A \subseteq \mathrm{N}^k$ is definable by a Σ_1-formula iff it is definable by a Σ-formula iff it is *recursively enumerable* (r.e.).

Since the set $\mathrm{Thm}(T)$ of all sentences provable in a recursively axiomatizable theory T is recursively enumerable, one can choose an arithmetical formula $\mathsf{Pr}_T(x) \in \Sigma_1$ which defines the set $\mathrm{Thm}(T)$ in $\mathbf{N}$, i.e. satisfies the equivalence $T \vdash \varphi \Leftrightarrow \mathbf{N} \models \mathsf{Pr}_T(\overline{\varphi})$ for each sentence φ in the language of T. The formula $\mathsf{Pr}_T(x)$ can be read the sentence x is provable in T; it is called a *provability predicate* of the theory T.

Peano arithmetic PA is a theory formulated in the arithmetical language; its axiom set consists of seven (eight, or nine, in some sources) single axioms and one schema, the *induction schema.* The Σ-*completeness theorem* says that each Σ-sentence valid in $\mathbf{N}$ is provable in PA. A simple consequence of the Σ-completeness theorem is the following. If $\mathsf{Pr}_{\mathrm{PA}}(x)$ is any Σ_1-formula which defines the set $\mathrm{Thm}(\mathrm{PA})$ of all sentences provable in PA then, besides the equivalence $\mathrm{PA} \vdash \varphi \Leftrightarrow \mathbf{N} \models \mathsf{Pr}_{\mathrm{PA}}(\overline{\varphi})$ for any arithmetical sentence φ, it also satisfies the implication $\mathrm{PA} \vdash \varphi \Rightarrow \mathrm{PA} \vdash \mathsf{Pr}_{\mathrm{PA}}(\overline{\varphi})$. This implication is known as the *first derivability condition D1*; it is a property shared by all provability predicates. There are of course many (true, i.e. valid in $\mathbf{N}$) sentences that are provable in PA but the provability of which is not obtained by an appeal to the Σ-completeness theorem. An example is the sentence for each x there is a prime number y greater than x; this sentence is not Σ.

The sentence $\neg\mathsf{Pr}_{\mathrm{PA}}(\overline{0 = \mathrm{S}(0)})$, saying that contradiction is not provable in PA, is denoted $\mathsf{Con}(\mathrm{PA})$ and called a *consistency statement* of PA (based on the formula $\mathsf{Pr}_{\mathrm{PA}}(x)$). Since $\mathrm{PA} \nvdash 0 = \mathrm{S}(0)$, from the fact that the formula $\mathsf{Pr}_{\mathrm{PA}}(x)$ defines the set $\mathrm{Thm}(\mathrm{PA})$ we have $\mathbf{N} \models \mathsf{Con}(\mathrm{PA})$.

Given a notion like being a power of two or being a (numerical code of a) sentence provable in PA, it is natural to ask the following question: if ψ is a formula that defines that notion in $\mathbf{N}$, can we take for granted that "canonical facts" about that notion, if expressed in the arithmetical language using the formula ψ, are provable in PA? The answer is no in general, but yes in many cases and if the formula ψ is chosen properly. A canonical fact about powers of two is that a number x is a power of two if and only if $x = \overline{1}$ or $x = \overline{2}\cdot y$ where y is a power of two. If powers of two are represented by the formula $\mathsf{Pow}(x)$ defined above then indeed, this fact is provable in PA. It is however not so difficult to find a formula $\psi(x)$ which also defines the set of all powers of two but does not have this additional property.

As to formalized provability, the canonical facts, i.e. the desired additional properties of the formula $\mathsf{Pr}_{\mathrm{PA}}(x)$, are the *second and third derivability conditions D2 and D3,* where D2 is $\mathrm{PA} \vdash \mathsf{Pr}_{\mathrm{PA}}(\overline{\varphi \to \psi}) \to (\mathsf{Pr}_{\mathrm{PA}}(\overline{\varphi}) \to \mathsf{Pr}_{\mathrm{PA}}(\overline{\psi}))$ and D3 is $\mathrm{PA} \vdash \mathsf{Pr}_{\mathrm{PA}}(\overline{\varphi}) \to \mathsf{Pr}_{\mathrm{PA}}(\overline{\mathsf{Pr}_{\mathrm{PA}}(\overline{\varphi})})$. With some effort one can show that the "normal" choice of the formula $\mathsf{Pr}_{\mathrm{PA}}(x)$ defining the set $\mathrm{Thm}(\mathrm{PA})$ satisfies the conditions D2 and D3 (i.e. satisfies all derivability conditions D1–D3) for all sentences φ and ψ. The condition D3 is in fact a consequence of a more general fact, namely the *formalized* Σ-*completeness theorem* saying that $\mathrm{PA} \vdash \sigma \to \mathsf{Pr}_{\mathrm{PA}}(\overline{\sigma})$ for each Σ-sentence σ. The last thing we need to

mention is the proof predicate. One can assume that the provability predicate $\mathsf{Pr}_{\mathrm{PA}}(x)$ has the form $\exists y \mathsf{Proof}_{\mathrm{PA}}(x,y)$ where $\mathsf{Proof}_{\mathrm{PA}}(x,y) \in \Delta_0$. The formula $\mathsf{Proof}_{\mathrm{PA}}(x,y)$ is a *proof predicate* of PA; it can be read the number y is a proof of the sentence x.

3 Meta-mathematics and self-reference

The *self-reference theorem* says that for any arithmetical formula $\psi(x)$ there exists a sentence φ such that $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$. The sentence φ satisfying $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$, i.e. provably equivalent to the sentence the sentence φ has the property ψ, can naturally be viewed as a first-person sentence, saying I have the property ψ. It is often convenient to view the expression $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$ as an *equation,* determined by the formula ψ, with an unknown sentence φ. So the self-reference theorem ensures that every self-referential equation has a solution.

A prominent example on the use of the self-reference theorem is *Gödel sentence* saying I am not provable in PA, i.e. a sentence ν satisfying the condition $\mathrm{PA} \vdash \nu \equiv \neg \mathsf{Pr}_{\mathrm{PA}}(\overline{\nu})$. A second important example is *Rosser sentence* ρ satisfying

$$\mathrm{PA} \vdash \rho \equiv \exists y (\mathsf{Proof}_{\mathrm{PA}}(\overline{\neg \rho}, y) \,\&\, \forall v \leq y \neg \mathsf{Proof}_{\mathrm{PA}}(\overline{\rho}, v)), \tag{1}$$

i.e. saying there is a proof of my negation such that beneath it there is no proof of myself. Both Gödel and Rosser sentences demonstrate incompleteness of Peano arithmetic. An important distinction between these two sentences is that while independence of Rosser sentence is formalizable in PA, viz $\mathrm{PA} \vdash \mathsf{Con}(\mathrm{PA}) \to \neg \mathsf{Pr}_{\mathrm{PA}}(\overline{\rho}) \,\&\, \neg \mathsf{Pr}_{\mathrm{PA}}(\overline{\neg \rho})$, only one half of independence of Gödel sentence can be formalized: $\mathrm{PA} \vdash \mathsf{Con}(\mathrm{PA}) \to \neg \mathsf{Pr}_{\mathrm{PA}}(\overline{\nu})$, but $\mathrm{PA} \nvdash \mathsf{Con}(\mathrm{PA}) \to \neg \mathsf{Pr}_{\mathrm{PA}}(\overline{\neg \nu})$. Since my point is that properties of self-referential sentences can be obtained using modal logic, no proofs are given here.

L. Henkin asked a question whether a sentence saying I am provable must be provable. In the equational setting, Henkin's question concerns the equation $\mathrm{PA} \vdash \kappa \equiv \mathsf{Pr}_{\mathrm{PA}}(\overline{\kappa})$ for an unknown sentence κ. Since this equation has a trivial (provable) solution $0 = 0$, Henkin's question should be understood as a question whether *any* Henkin sentence is provable, i.e. whether the Henkin equation has, up to provable equivalence, only one solution. The question brought the problem of uniqueness of self-referential equations to the attention of logicians and was answered positively in the influential paper by Löb (1955).

Let $\mathrm{PA}\restriction y + x$ denote the theory whose axioms are the sentence x plus all axioms of PA which are less than y. Our fourth example of a self-referential

equation is

$$\mathrm{PA} \vdash \eta \equiv \exists y(\neg \mathsf{Con}(\mathrm{PA}\restriction y + \overline{\eta}) \,\&\, \mathsf{Con}(\mathrm{PA}\restriction y + \overline{\neg\eta})). \tag{2}$$

Any sentence η satisfying this equation has the property that both theories $\mathrm{PA}+\eta$ and $\mathrm{PA}+\neg\eta$ are *interpretable* in PA, in the syntactical sense of Tarski, Mostowski, and Robinson (1953). This construction of a symmetrically interpretable sentence is mentioned e.g. in Švejdar (1983), but is in fact distilled out from Hájek and Hájková (1972); so the sentence η could be called *Hájek sentence*. However it should be noted that other self-referential constructions named by Petr Hájek exist.

4 Self-reference and modal logic

A crucial idea, connecting meta-mathematics and modal logic, is interpreting the usual modal symbol $\Box$, the necessity operator, by formalized provability expressed by the formula $\mathsf{Pr}_{\mathrm{PA}}(x)$. We directly proceed to modal systems with two additional "modalities" that are not met in traditional modal logical studies and that are closely related one to another, the Rosser symbols $\preceq, \prec$.

So we consider propositional modal language with propositional atoms, the symbol $\bot$ for falsity, connectives $\to$, $\&$, $\vee$, the unary modality $\Box$, and two binary *Rosser* (or *witness comparison*) modalities $\preceq$ and $\prec$. *Modal formulas* are built up from atoms and $\bot$ using connectives and modalities, with the restriction that $\preceq, \prec$ are applicable only to formulas starting with $\Box$. So $\Box(\Box p \to p)$ or $q \,\&\, (\Box p \preceq \Box\bot)$ are examples of modal formulas. We use $\neg A$, $\top$, $A \equiv B$, and $\Diamond A$ as shorthands for $A \to \bot$, $\bot \to \bot$, $(A \to B) \,\&\, (B \to A)$, and $\neg\Box\neg A$ respectively. For omitting parentheses, we assign the symbols $\preceq, \prec$ higher priority than binary connectives; among connectives, implication $\to$ has higher priority than equivalence $\equiv$, but lower that conjunction $\&$ and disjunction $\vee$. So $p \equiv q \vee \Box p \prec \Box q$ is the same formula as $p \equiv (q \vee (\Box p \prec \Box q))$.

To define the *arithmetical semantics* for modal formulas with Rosser modalities we need two auxiliary notions, standard proof predicate and arithmetical evaluation. A *standard proof predicate* is an arithmetical formula $Prf(x, y)$ satisfying $\mathrm{PA} \vdash \forall x(\exists y Prf(x, y) \equiv \mathsf{Pr}_{\mathrm{PA}}(x))$ and such that both $Prf(x, y)$ and $\neg Prf(x, y)$ are PA-equivalent to a Σ_1-formula (i.e., $Prf(x, y)$ is a $\Delta_1(\mathrm{PA})$-formula). The formula $\mathsf{Proof}_{\mathrm{PA}}(x, y)$ is an example of a standard proof predicate. It is evident that if $Prf(x, y)$ is a standard proof predicate then the formula $\exists y Prf(x, y)$, which may be called a provability predicate associated with that proof predicate, satisfies the derivability conditions D1–D3. A function e from modal formulas to arithmetical sentences is an *(arithmetical) evaluation* based on a standard proof predicate $Prf(x, y)$ if it commutes with logical connectives ($e(A \& B) = e(A) \& e(B)$, etc.) and satisfies

$e(\bot) = (0 = \mathrm{S}(0))$ and

$$\begin{aligned} e(\Box A) &= \exists y Prf(\overline{e(A)}, y), \\ e(\Box A \preceq \Box B) &= \exists y (Prf(\overline{e(A)}, y) \,\&\, \forall v{<}y \neg Prf(\overline{e(B)}, v)), \\ e(\Box A \prec \Box B) &= \exists y (Prf(\overline{e(A)}, y) \,\&\, \forall v{\leq}y \neg Prf(\overline{e(B)}, v)). \end{aligned}$$

Thus the modal formulas $\Box A \preceq \Box B$ and $\Box A \prec \Box B$ can be read "A has a proof which is less than or equal to (or less than, respectively) any possible proof of B". It is natural to read the formula $\Box A$ as "A is provable" rather than "A is necessary".

A modal formula is a PA-*tautology* if $\mathrm{PA} \vdash e(A)$ for each evaluation e based on any standard proof predicate $Prf(x, y)$.

Note that an evaluation is fully determined by its values on propositional atoms and that a value $e(A)$ of a formula A depends on only those atoms that occur in A. So if A is a formula like $\neg\Box\bot$, containing no atoms, $e(A)$ is the same sentence for each e. Also note that if A contains no occurrences of $\preceq, \prec$ then $e(A)$ is equivalent to a sentence constructed from values of atoms using only logical connectives and the formula $\mathsf{Pr}_{\mathrm{PA}}(x)$; in this sense $e(A)$ is independent of the choice of the proof predicate $Prf(x, y)$. This follows from the fact that if $\Box B$ is any subformula of A then $e(\Box B)$ is PA-equivalent to $\mathsf{Pr}_{\mathrm{PA}}(\overline{e(B)})$.

The formula $\Box(\bot \to p)$ is an example of a PA-tautology because PA knows (can prove) that ex falso quodlibet. The formula $\neg\Box\bot$ is *not* a PA-tautology because its value $e(A)$ is (for each e) the sentence $\mathsf{Con}(\mathrm{PA})$ which, by Gödel Second Incompleteness Theorem, is not provable in PA. Now recall that $e(\bot)$ is $0 = \mathrm{S}(0)$ for each e, and that $e(\top)$ is provable and $e(\bot)$ is refutable in PA. So if e is based on a standard proof predicate $Prf(x, y)$ then $\mathbf{N} \models \exists y Prf(\overline{e(\top)}, y)$ and $\mathbf{N} \not\models \exists y Prf(\overline{e(\bot)}, y)$. Hence $\mathbf{N} \models Prf(\overline{e(\top)}, \overline{m})$ for some $\overline{m}$ and $\mathbf{N} \models \neg Prf(\overline{e(\bot)}, \overline{k})$ for each k. Since Σ-completeness is applicable to both formulas Prf and $\neg Prf$, we have $\mathrm{PA} \vdash Prf(\overline{e(\top)}, \overline{m})$ and $\mathrm{PA} \vdash \forall v{\leq}\overline{m} \neg Prf(\overline{e(\bot)}, v)$. Thus

$$\mathrm{PA} \vdash \exists y (Prf(\overline{e(\top)}, y) \,\&\, \forall v{\leq}y \neg Prf(\overline{e(\bot)}, v)).$$

This argument shows that $\Box\top \prec \Box\bot$ is another example of a PA-tautology; PA knows that the trivially true sentence $\neg(0 = \mathrm{S}(0))$ has a proof less than any proof of contradiction without discussing the *existence* of a proof of contradiction. Now consider, as our final example, the two trivially provable sentences $\neg(0 = \mathrm{S}(0))$ and $\mathsf{Pr}_{\mathrm{PA}}(\overline{\neg(0 = \mathrm{S}(0))})$ which are the values $e(\top)$ and $e(\Box\top)$ of the modal formulas $\top$ and $\Box\top$ under any evaluation e. It is possible to choose a standard proof predicate $Prf(x, y)$ and an evaluation e based on it such that, in the sense of $Prf(x, y)$, the first proof of $e(\Box\top)$ is less than the first proof of $e(\top)$. Then $\mathrm{PA} \not\vdash e(\Box\top \preceq \Box\Box\top)$ and so $\Box\top \preceq \Box\Box\top$ is not a

PA-tautology. The same argument together with a choice of another proof predicate shows that neither $\Box\Box\top \prec \Box\top$ is a PA-tautology. This example demonstrates why the notion of standard proof predicate is useful: letting the proof predicate vary, i.e. opening the possibility of artificial reordering proofs, is a means how to get rid of irrelevant facts (possibly depending on details of coding of syntax) about ordering of proofs.

The *modal logic* R of Guaspari and Solovay (1979), the classical theory of witness comparison, has the following axiom schemas A1–A4, B1–B4, and P and deduction rules MP, Nec, and Un:

A1: all propositional tautologies,

A2: $\Box(A \to B) \to (\Box A \to \Box B)$,

A3: $\Box A \to \Box\Box A$,

A4: $\Box(\Box A \to A) \to \Box A$,

MP: $A,\ A \to B\ /\ B$, Nec: $A\ /\ \Box A$,

Un: $\Box A\ /\ A$,

B1: $\Box A \preceq \Box B \to \Box A$,

B2: $\Box A \preceq \Box B\ \&\ \Box B \preceq \Box C \to \Box A \preceq \Box C$,

B3: $\Box A \vee \Box B \to \Box A \preceq \Box B \vee \Box B \prec \Box A$,

B4: $\Box A \prec \Box B \equiv \Box A \preceq \Box B\ \&\ \neg(\Box B \preceq \Box A)$,

P: $\Box A \preceq \Box B \to \Box(\Box A \preceq \Box B)$, $\Box A \prec \Box B \to \Box(\Box A \prec \Box B)$.

The restriction of the logic R to the usual modal language, with $\Box$ as the only modality, axioms A1–A4, and rules modus ponens MP and necessitation Nec, is *provability logic* GL. Nowadays, there is a rich literature about provability logic. The logic R as well as the other systems mentioned below are conservative extensions of provability logic. There is no need to add the unnecessitation rule Un to provability logic because Un is a derived rule in that logic. The schema P could be called a *persistency schema.* We keep it a little bit apart from the *basic axioms B1–B4* about witness comparison because we will discuss some alternatives to it.

The logic R is complete with respect to the arithmetical semantics. It also has a satisfactory Kripke semantics and is decidable; which is quite surprising in the light of the highly inefficient nature of the arithmetical semantics. Out of these facts, it is rather straightforward to verify soundness of all axioms and rules, perhaps with an exception of A4, the Löb axiom schema. Note that soundness of Nec, A2, and A3 follows immediately from the derivability conditions, while soundness of the schema P follows from the formalized Σ-completeness theorem.

Consider the modal formula

$$\Box(p \equiv \Box\neg p \prec \Box p) \to (\neg\Box\bot \to \neg\Box p\ \&\ \neg\Box\neg p). \tag{3}$$

It says that "if p is provably equivalent to the statement saying that there is a proof of $\neg p$ which is less than any proof of p, then, under the assumption of consistency, neither p nor $\neg p$ is provable". Formula (3) is an example of a formula provable in R; this fact is left as an exercise because we give a similar proof in our logic SR below. Then, by arithmetical soundness of the logic R, the arithmetical value of formula (3) is provable in PA under any evaluation. Let e be an evaluation sending p to the Rosser sentence ρ. For this evaluation we also have $\mathrm{PA} \vdash e(\Box(p \equiv \Box\neg p \prec \Box p))$ and so $\mathrm{PA} \vdash e(\neg\Box\bot \to \neg\Box p \,\&\, \neg\Box\neg p)$. Since $e(\neg\Box\bot \to \neg\Box p \,\&\, \neg\Box\neg p)$ is the sentence $\mathsf{Con}(\mathrm{PA}) \to \neg\mathsf{Pr}_{\mathrm{PA}}(\overline{\rho}) \,\&\, \neg\mathsf{Pr}_{\mathrm{PA}}(\overline{\neg\rho})$, we have proved the fact mentioned above when speaking about Gödel and Rosser sentences: independence of the Rosser sentence ρ is formalizable in PA.

The reasoning in the previous paragraph is an example of a "direct" application of modal logic: properties of a particular self-referential sentence can be obtained by proving appropriate modal formulas in an appropriate modal system. Here is another example of a direct application of modal logic in meta-mathematics: from provability of the modal formula $\Box(p \equiv \neg\Box p) \to (\neg\Box\bot \to \neg\Box p)$ in R (and in GL) we get that PA proves the sentence $\mathsf{Con}(\mathrm{PA}) \to \neg\mathsf{Pr}_{\mathrm{PA}}(\overline{\nu})$ expressing formalized unprovability of the Gödel sentence ν. An example of a very nice and less direct application of modal logic is this: a sentence like ρ, whose full independence (unprovability and unrefutability) is formalizable in PA, cannot be obtained by *Gödelian self-reference,* i.e. by writing down an equation $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$ with ψ speaking about provability only, not employing things like the Rosser trick. Out of our examples, Gödel and Henkin sentences are obtained by Gödelian self-reference, Rosser and Hájek sentences are not.

5 Alternative systems with Rosser modalities

The Hájek sentence (2) says that a number y such that $\neg\mathsf{Con}(\mathrm{PA}\restriction y + \overline{\eta})$ appears before any possible v such that $\neg\mathsf{Con}(\mathrm{PA}\restriction v + \overline{\neg\eta})$; so it in fact also uses the Rosser trick. Moreover, the formula $\exists y \neg\mathsf{Con}(\mathrm{PA}\restriction y + \neg x)$ is equivalent to $\mathsf{Pr}_{\mathrm{PA}}(x)$; so it can be viewed as a *generalized proof predicate*, and a y such that $\neg\mathsf{Con}(\mathrm{PA}\restriction y + \neg x)$ can be viewed as a proof of x in a generalized sense. Put together, the Hájek sentence is a sort of a Rosser sentence. However, the modal system R does not appy to it because the sentence is not a Σ-sentence. Speaking otherwise, if evaluations e based on the generalized proof predicate are allowed, then the formalized Σ-completeness theorem is not applicable to sentences of the form $e(\Box A \preceq \Box B)$ and $e(\Box A \prec \Box B)$, and thus the schema P is not arithmetically sound.

A modal system applicable to the Hájek sentence was proposed in Švejdar (1983). The system is denoted Z there; it results from replacing the schema P

by another schema Z:

Z: $\quad \Box A \to \Box(\neg B \to \Box A \prec \Box B).$

This schema says that "if A is provable then it can be proved that it has a proof smaller than any proof of any false statement B", or more briefly, "false statements do not have small proofs". It is not difficult to verify that formula (3), the modalized Rosser theorem, is provable also in the logic Z and that the schema Z is provable in the logic R. Soundness of the schema Z w.r.t semantics where evaluations can be based also on the generalized proof predicate $\neg\mathsf{Con}(\mathrm{PA}\restriction y + \neg x)$ follows from Theorem 1 below. So (formalized or not) independence of the Hájek sentence is in fact a consequence of modal considerations. It is also known (Švejdar, 1983) that the logic Z has a Kripke completeness theorem and is decidable.

Independence is not the only important property of the Hájek sentence by far. But since the system Z cannot prove any formula which is not provable in R, it offers no modal explanation why the Hájek sentence (2) should be important. The goal of this paper is to suggest a system stronger than Z that could express what Hájek sentence and similar constructions have in comparison to the usual Rosser sentence.

Let a *(modal)* Σ*-formula* be any formula obtained from formulas starting with $\Box$ and from $\bot$ and $\top$ using only conjunctions and disjunctions. Let S be the schema

S: $\quad \Box(E \to A) \to \Box(E \,\&\, \neg B \to \Box A \prec \Box B),$

where E is a Σ-formula and A and B are arbitrary modal formulas. Let SR be the system like Z, but with the schema Z replaced by the schema S. The notation is somewhat tentative: The letter "S" refers to Greek Σ because the logic SR is a Rosserian system with a special attention to arithmetical Σ-sentences, or "SR" can mean "semi-reflexive".

Theorem 1 *The logic* SR *is sound w.r.t. the arithmetical semantics with evaluations based on the generalized proof predicate* $\neg\mathsf{Con}(\mathrm{PA}\restriction y + \neg x)$.

Proof We show arithmetical soundness of the schema S. A key step is to use the *essential reflexivity* of PA: the implication $\chi \to \mathsf{Con}(\mathrm{PA}\restriction \overline{k} + \overline{\chi})$ is provable in PA for each m and each sentence χ. So let modal formulas A and B, modal Σ-formula E and an evaluation e based on the generalized proof predicate be given. Let φ, ψ, and σ be the arithmetical sentences that are the values of formulas A, B, and E (respectively) under v. We have to prove in PA that if $\sigma \to \varphi$ is provable then one can prove, using the assumptions σ and $\neg\psi$, that $\overline{\varphi}$ has a generalized proof smaller than any generalized proof of $\overline{\psi}$. We shift the speech one level up, giving a meta-mathematical argument. The

real proof of the theorem is then obtained by formalizing the considerations in PA, i.e. essentially by rewriting the argument using the sans-serif font. So let $\mathrm{PA} \vdash \sigma \to \varphi$. We have to show

$$\mathrm{PA} \vdash \sigma \,\&\, \neg\psi \to \exists y(\neg\mathsf{Con}(\mathrm{PA}\restriction y + \overline{\neg\varphi}) \,\&\, \forall v{\leq}y\,\mathsf{Con}(\mathrm{PA}\restriction v + \overline{\neg\psi})). \tag{4}$$

Note that $\forall v{\leq}y\,\mathsf{Con}(\mathrm{PA}\restriction v + \overline{\neg\psi})$ is the same as $\mathsf{Con}(\mathrm{PA}\restriction y + \overline{\neg\psi})$. A fixed proof of $\sigma \to \varphi$ uses only a finite number of axioms of PA. So if m_1 is sufficiently big then the axioms of PA less than m_1, together with the sentence $\neg(\sigma{\to}\varphi)$, constitute a contradictory theory. Thus, in fact by Σ-completeness, we have $\mathrm{PA} \vdash \neg\mathsf{Con}(\mathrm{PA}\restriction \overline{m_1} + \overline{\neg(\sigma \to \varphi)})$. An inspection of the proof of formalized Σ-completeness yields an m_0 such that $\mathrm{PA} \vdash \sigma{\to}\neg\mathsf{Con}(\mathrm{PA}\restriction \overline{m_0}{+}\overline{\neg\sigma})$ for (each and thus) our sentence σ. So if $m = \max\{m_0, m_1\}$ we have

$$\mathrm{PA} \vdash \sigma \to \neg\mathsf{Con}(\mathrm{PA}\restriction \overline{m} + \overline{\neg\varphi}), \tag{5}$$

m is a generalized proof of φ. Essential reflexivity applied to the sentence ψ yields $\mathrm{PA} \vdash \neg\psi \to \mathsf{Con}(\mathrm{PA}\restriction \overline{m} + \overline{\neg\psi})$. This and (5) implies (4), q.e.d. ∎

Note the difference between the proof predicate $\neg\mathsf{Con}(\mathrm{PA}\restriction y + \neg x)$ and the standard proof predicate $\mathsf{Proof}_{\mathrm{PA}}(x, y)$: there is no common m_0 such that $\sigma \to \exists v{\leq}\overline{m_0}\,\mathsf{Proof}_{\mathrm{PA}}(\overline{\sigma}, v)$ is provable for each Σ-sentence σ.

A substitution of $\Box\bot$, $\bot$, and $\Box\bot$ for A, B, and E in the schema S yields $\Box(\Box\bot \to \Box\Box\bot \prec \Box\bot)$. Then, using the rule Un, we get $\Box\bot \to \Box\Box\bot \prec \Box\bot$. This formula, as well as the formula $\Box\Box\bot \to \Box\Box\bot \prec \Box\bot$, are examples of formulas provable in the logic SR but unprovable in R. As an exercise, the reader may try to derive the latter from the former by distinguishing cases $\Box\bot$ and $\neg\Box\bot$. A further example of a formula provable in SR is

$$\Box(p \equiv \Box\neg p \prec \Box p) \to (\Box(E \to p) \vee \Box(E \to \neg p) \to \Box\neg E), \tag{6}$$

for any Σ-formula E. For, let D be $\Box(p \equiv \Box\neg p \prec \Box p)$. Then the proof of (6) in SR may proceed like this:

$$\begin{aligned}
&\Box(E \to p) \to \Box(E \,\&\, p \to \Box p \prec \Box\neg p) && ;\ \mathrm{S}\\
&\Box(E \to p) \to \Box(E \to \Box p \prec \Box\neg p)\\
&\Box(E \to p) \to \Box(E \to \Box p \preceq \Box\neg p) && ;\ \mathrm{B4}\\
&\Box(E \to p) \to \Box(E \to \neg(\Box\neg p \prec \Box p)) && ;\ \mathrm{B4}\\
&D \to \Box(p \to \Box\neg p \prec \Box p)\\
D \,\&\, &\Box(E \to p) \to \Box(E \to \Box\neg p \prec \Box p)\\
D \,\&\, &\Box(E \to p) \to \Box\neg E,
\end{aligned}$$

where the seventh line follows from the sixth and fourth. The proof of the second half, $D \,\&\, \Box(E \to \neg p) \to \Box\neg E$, is similar.

A substitution of $\top$ for E in (6) yields formula (3); so (6) is a stronger version of modalized Rosser theorem. Formula (6), as well as formula (3), speaks about properties of the Rosser sentence, i.e. sentence p provably equivalent to the statement there is a proof of $\neg p$ which is less than any proof of p. While formula (3) says that the Rosser sentence is independent provided the base theory is consistent, formula (6) says that neither p nor $\neg p$ can be proved using any assumption expressed by a consistent Σ-sentence. It is known that, in PA, the unprovability of a sentence φ from any consistent Σ-sentence is the same as interpretability of PA + $\neg\varphi$ in PA. This fact is due to Petr Hájek. So provability of formula (6) in SR together with Theorem 1 yields the symmetric interpretability of the Hájek sentence.

Speaking precisely, there is a little gap in the conclusion that the symmetric interpretability of Hájek sentence can be obtained modally. Our considerations only show that neither Hájek sentence η nor its negation $\neg\eta$ can be proved from a consistent assumption *which is a value of a modal* Σ*-formula.* Note that values of modal Σ-formulas constitute a proper subset of all Σ-sentences. To get the full result, one would modify the logic SR as follows: (i) consider two sorts of propositional atoms, (normal) atoms p, q, ... and Σ-atoms σ, τ, ..., and (ii) prove a formula like (6), but with E replaced by a Σ-atom σ. We did not do this modification to keep things simpler.

To finish, it must be admitted that we are still rather far from a well-rounded modal system for Rosser modalities. Nevertheless, I believe that the schemas Z and S do formalize arguments that are quite frequent when dealing with self-referential sentences that use some version of the Rosser trick.

Vítězslav Švejdar
Department of Logic, Charles University
Palachovo nám. 2, 116 38 Praha 1, Czech Republic
vitezslavdotsvejdaratcunidotcz, http://www1.cuni.cz/~svejdar/

References

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, *37*, 349–360.

Guaspari, D., & Solovay, R. M. (1979). Rosser sentences. *Annals of Math. Logic*, *16*, 81–99.

Hájek, P., & Hájková, M. (1972). On interpretability in theories containing arithmetic. *Fundamenta Mathematicae*, *76*, 131–137.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *J. Symbolic Logic*, *20*, 115–118.

Smoryński, C. (1985). *Self-reference and modal logic.* New York: Springer.

Solovay, R. M. (1976). Provability interpretations of modal logic. *Israel J. Math.*, *25*, 287–304.

Švejdar, V. (1983). Modal analysis of generalized Rosser sentences. *J. Symbolic Logic*, *48*(4), 986–999.

Tarski, A., Mostowski, A., & Robinson, R. M. (1953). *Undecidable theories*. Amsterdam: North-Holland.