A note on applicability of the incompleteness theorem to human mind

Pavel Pudlák

Mathematical Institute, Academy of Sciences of Czech Republic

Abstract

We shall present some relations between consistency and reflection principles which explain why is Gödel's incompleteness theorem wrongly used to argue that thinking machines are impossible.

1 Introduction

Since its publishing Gödel's incompleteness theorem attracted a lot of attention among philosophers. In 1959 Lucas [8] presented an argument that this theorem implies that human thinking is essentially different from what any machine can do. This means that the ultimate goal of *artificial intelligence* cannot be achieved. The argument is roughly the following. A machine (nowadays we would rather say "a computer") behaves according to fixed rules (a program), hence we can view it as a formal system. Applying Gödel's theorem to this system we get a true sentence which is unprovable in the system. Thus the machine does not know that the sentence is true while we can see that it is true.

The spectrum of attitudes of various people to this argument was nicely characterized by Hofstadter [5, page 472]: *"Some size onto it as a nearly religious proof of the existence of souls, while others laugh it off as being unworthy of comment"*. Lucas's argument has been criticized several times. In particular, in his famous book [5] Hofstadter analyzed it in details and gave several founded counterarguments. Still in 1989 and 1994 Penrose published two books [9,10]

[1] The author was partially supported by grant A1019602 of the Academy of Sciences of the Czech Republic and by a joint grant of NSF (USA) and MŠMT (Czech Rep.) INT-9600919/ME-103.

where he defended the thesis of Lucas.[2] He went even further and concluded that there must be some physical phenomena that our brains make use of and which we do not know yet. Especially in the second book he analyzed the argument in details and took pains to consider and dismiss many possible counterarguments.

It seems that most logicians agree on that Gödel's theorem is not relevant to the question whether an intelligent machine can be constructed. Curiously enough, they do not agree so well on what is wrong with Lucas's argument. In his crushing review of the second Penrose's book, Putnam [12] argues, loosely interpreting his argument, that a computer program simulating human intelligence must be extremely complex, thus we cannot completely appreciate it and produce a true independent sentence. However, one can argue that already some present programs and chips are extremely complex and nobody can be sure that they do not contain a serious bug. Still we (more precisely those who designed them) know what they were intended for. Then, assuming that they were produced correctly and they are supposed to prove sentences, we can easily produce an unprovable true sentence.

The main argument of Hofstadter is based on the distinction between the system to which we apply Gödel's theorem and the system in which we perform the argument. This is the basic distinction between a theory and metatheory in logic, which is inevitable, if we do not want to run into trivial inconsistencies. A person reasoning about a machine knows the machine completely, thus there is nothing surprising in being able to produce something that the machine cannot prove. This is completely symmetric with respect to interchanging the roles of the mind and the machine, therefore we cannot conclude that they are different. It is not possible to produce such statements, however, if a subject reasons about itself.

Still, our personal experience seems to suggest that we can somehow "step out" and avoid Gödel's theorem. In order to explain this, one could refer to the tremendous complexity of the human mind, to its inconsistency, vagueness and possibly other deficiencies. But let us consider just mathematical thinking. Actually Gödel's sentence is not just some nonsensical statement, even if it is constructed for a very complex system. It expresses the consistency of the system, which is a clear mathematical statement. In mathematics people do also a lot of mistakes, but, in principle, their mathematical reasoning is exact. Thus vagueness and inconsistency of human thinking does not explain it. The argument using complexity can be rejected as well, since the consistency of the system depends only on the mathematical assumptions that people use, not on the amount and complexity of the results they use. As far as the basic assumptions are concerned, almost all mathematicians use just a part of ZFC

[2] Recently he published another one [11] which I had not chance to look at.

and all the axioms of ZFC (more precisely, axiom schemas) can be written on a single page. To produce an independent sentence for a human mind or a computer, we do not have to analyze it in its whole complexity, we only need to know the set theoretical assumptions that it uses.

This shows that the relevant question is what mathematical statements we are willing to accept as intuitively true. This question has been considered by many logicians. It has been studied quite formally in proof theory and systems which should capture our mathematical assumptions have been proposed, [2,3,6]. Naturally, the incompleteness phenomenon plays a key role also there. Hilbert [4] proposed to develop all mathematics using only *finite means*. It is very difficult to characterize such *finitism*. Kreisel [6] argued that *"if the notion of finitist proof is capable of formalization at all, its proof predicate must be not recognizable as such by finitist means"*. Let us note that the argument demonstrating this thesis refers to Gödel's theorem in a similar same way as Lucas's.

In this note I will concentrate on the phenomenon, or rather illusion, that we can always extend our assumptions by a true independent sentence. This is something that should be carefully analysed independently on ones attitude to Lucas type arguments. (If, for instance, one believes that human mind is superior to any artificial device, because of some unknown phenomena, one can get at least some hints about the new phenomena in this way.) My explanation will be very simple: when arguing that the new system is consistent, we use unconsciously a stronger assumption. Still, I think that writing down explicitly the assumptions and relationships between them will help to clarify the subject.

2 Preliminaries

Our base theory, denoted by B, will be $I\Sigma_1$. This particular choice of a theory is not essential, one can take, for instance PA (Peano arithmetic), or $I\Sigma_0+Exp$ (with some arguments slightly modified). If not stated otherwise, all theorems are claimed to be provable in B.

By a *theory* we mean any recursively axiomatizable set of sentences in some language. In this paper we shall consider only extensions of B. If a theory is given by an infinite set of axioms, the way it is presented may influence provability of its consistency etc. Thus to be quite precise we shall identify a theory with an index of a recursively enumerable set.

For a natural number n, we denote by $\underline{n}$ the *numeral* n, i.e. a suitable closed term representing n; the standard approach is to take the term $S^n(0)$, where S is the successor function. The gödel number of a formula φ will be denoted by $\ulcorner\varphi\urcorner$; for a formula φ with a free variable x, we denote by $\ulcorner\varphi(\dot{x})\urcorner$ the gödel number of φ with the free variable replaced by the numeral representing x. This is a formalization of the function $n \mapsto$ *gödel number of* $\varphi(\underline{n})$. This function cannot be expressed by an arithmetical term in the usual language of arithmetic, but, for sake of simplicity of notation, we shall use it in formulas as a term. $\perp$ denotes a suitable contradiction, say $0 = 1$. As usual, $T + \varphi$ denotes the theory T extended by the axiom φ.

We shall denote by $Prf_T(x, y)$ a natural formalization of the relation *"x is a proof of y in T"*. $Pr_T(y)$ denotes $\exists x\ Prf_T(x, y)$, i.e. the *provability predicate* of T. The naturalness means that the fact that the proof predicate is closed under logical rules can be proved in B. In order to reduce the number of parentheses we shall abbreviate the formula $Pr_T(\ulcorner\varphi\urcorner)$ by $Pr_T\ulcorner\varphi\urcorner$. The formalization of the *consistency* of T will be denoted by Con_T, it is the formula $\neg Pr_T\ulcorner\perp\urcorner$.

The *Rosser sentence* Ro_T for T is the negation of a sentence ρ obtained by the following diagonalization

$$\rho \equiv \exists x(Prf_T(x, \ulcorner\neg\rho\urcorner) \wedge \forall y < x\ \neg Prf_T(y, \ulcorner\rho\urcorner)).$$

(Sometimes ρ itself is called the Rosser sentence.)

ω-consistency of T, denoted by ω-$\mathbf{Con}_T$, is the schema (therefore we use boldface letters)

$$Pr_T\ulcorner\exists x\ \varphi(x)\urcorner \to \exists x\ \neg Pr_T\ulcorner\neg\varphi(\dot{x})\urcorner$$

for every formula $\varphi(x)$ with only x free. This schema restricted to primitive recursive formulas is 1-*consistency* and it will be denoted by $\mathbf{1}$-$\mathbf{Con}_T$. Here we shall identify primitive recursive formulas with those which are Δ_1 provably in B.

The *reflection principle* for T, denoted by $\mathbf{Rfn}_T$, is the schema $Pr_T\ulcorner\varphi\urcorner \to \varphi$ for every sentence φ in the language of T. The reflection principle for T restricted to a class of sentences Γ will be denoted by denoted by Γ-$\mathbf{Rfn}_T$. The *uniform reflection principle* for T, denoted by $\mathbf{RFN}_T$, is the schema $\forall x\ (Pr_T\ulcorner\varphi(\dot{x})\urcorner \to \varphi(x))$ for every sentence φ in the language of T. The uniform reflection principles restricted to classes of formulas Σ_n and Π_n are equivalent to sentences (namely, the uniform reflection for the corresponding universal formula), therefore they will be treated as such.

We shall use some well-known results on reflection principles.

Lemma 1 (see [13]) *(1) Σ_1-completeness of B: $\varphi \to Pr_B\ulcorner\varphi\urcorner$, for every Σ_1 sentence φ;*

(2) $Con_T \;\equiv\; \Pi_1\text{-}\mathbf{Rfn}_T \;\equiv\; \Pi_1\text{-}RFN_T$*;*

(3) $\mathbf{1\text{-}Con}_T \;\equiv\; \Sigma_1\text{-}\mathbf{Rfn}_T$*.*

Let us recall that Gödel's first incompleteness theorem asserts that for every ω-consistent T, there is an independent sentence. The sentence γ is defined by $\gamma \equiv \neg Pr_T\lceil\gamma\rceil$. The ω-consistency is needed only to prove that $\neg\gamma$ is unprovable, and, in fact, one needs only 1-consistency, while for unprovability of γ the plain consistency suffices. The second incompleteness theorem extends this by showing that γ is equivalent to Con_T. This can be expressed formally by

$$Con_T \rightarrow Con_{T+\neg Con_T}; \tag{1}$$

$$\mathbf{1\text{-}Con}_T \;\vdash\; Con_{T+Con_T}. \tag{2}$$

Rosser sentence for T is clearly implied by γ, thus (1) implies that it is not provable in T assuming T is consistent. Moreover $\neg Ro_T$ is unprovable using only Con_T; formally

$$Con_T \rightarrow Con_{T+Ro_T}. \tag{3}$$

Note that Rosser sentence, the consistency and Σ_1 reflection are of increasing strength and this hierarchy can be extended by taking Σ_n reflection schemas for $n = 2, 3, \ldots$.

3 The illusion of perpetual adding consistency

Gödel's theorem implies that the rule

$$\text{from } Con_T \text{ deduce } Con_{T+Con_T} \tag{4}$$

is not consistent with any sufficiently strong theory S. More precisely, we need that S proves the consistency of some theory for which it proves Gödel's theorem . This is true, say, for $I\Sigma_1$, which proves $Con_{I\Sigma_0+Exp}$, but also for weaker theories. To prove this claim, let T_0 be a theory for which we have $S \vdash Con_{T_0}$. Take $T = T_0 + \neg Con_{T_0}$; it is consistent by Gödel's theorem . But $T + Con_T$ is not consistent, since T proves that a subtheory of T is not consistent. Let us stress that (4) has the weakest possible form, since we apply it only for explicit theories.

By the same argument such a rule is inconsistent also for stronger sentences. (Note, however, that it does not make sense for schemas, such as $\mathbf{Rfn}_T$, since the corresponding rule would have infinitely many assumptions.) Namely, in case of Con we have derived a contradiction using (1). To prove that the above rule is contradictory for Σ_n-RFN_T, we shall check that the corresponding sentence is true for Σ_n-RFN_T.

Lemma 2 $\Sigma_n\text{-}RFN_T \ \rightarrow \ \Sigma_n\text{-}RFN_{T+\neg\Sigma_n\text{-}RFN_T}$.

Proof. Assume Σ_n-RFN_T. We need to prove, for a Σ_n formula $\varphi(x)$,

$$\forall x\ (Pr_T\ulcorner\neg\Sigma_n\text{-}RFN_T \rightarrow \varphi(\dot{x})\urcorner \rightarrow \varphi(x)).$$

Let x be given. Instead of $Pr_T\ulcorner\neg\Sigma_n\text{-}RFN_T \rightarrow \varphi(\dot{x})\urcorner$ we shall use a weaker assumption by taking only one special case of Σ_n-RFN_T, namely

$$Pr_T\ulcorner\neg(Pr_T\ulcorner\varphi(\dot{x})\urcorner \rightarrow \varphi(\dot{x})) \rightarrow \varphi(\dot{x})\urcorner.$$

The formula within the outer $\ulcorner\ \urcorner$ reduces (using propositional calculus) so that we get

$$Pr_T\ulcorner Pr_T\ulcorner\varphi(\dot{x})\urcorner \rightarrow \varphi(\dot{x})\urcorner.$$

By Löb's theorem (cf. [13]), it implies $Pr_T\ulcorner\varphi(\dot{x})\urcorner$. Now we can apply our assumption Σ_n-RFN_T and conclude $\varphi(x)$ as required. $\square$

Let us analyze now the intuitive argument that we can add the consistency Con_{T+Con_T} when we already know Con_T. The usual argument goes roughly as follows:

> Suppose $\neg Con_{T+Con_T}$, i.e. T proves $\neg Con_T$. Since T is *true*, there is an actual proof of contradiction from the axioms T. But this is in contradiction with our assumption Con_T.

It is clear that this argument uses an additional assumption that T is "true", whatever it means. This word appears in Lucas's argument [8, p. 117], while Penrose uses "sound" [10, pp. 75,94]. The way these words are used shows that their meaning is some version of the reflection principle for T. The weakest reflection principle (of those we have considered) which suffices for this argument is Σ_1-$\mathbf{Rfn}_T$. Since Σ_1-$\mathbf{Rfn}_T$ is equivalent to $\mathbf{1\text{-}Con_T}$, the argument is simply showing a half of Gödel's theorem , as expressed by the formula (2).

The fact (2) does not suffice to continue with adding more and more consistencies, but it is not difficult to prove a slightly stronger statement (we shall prove it formally below)

$$\mathbf{1\text{-}Con}_T \ \vdash \ \mathbf{1\text{-}Con}_{T+Con_T}. \tag{5}$$

This explains the illusion that we have the power to add consistencies for ever. If we start with some theory T_0 and assume $\mathbf{1\text{-}Con}_{T_0}$, then we can prove Con_{T_0}, $Con_{T_0+Con_{T_0}}$, $Con_{T_0+Con_{T_0+Con_{T_0}}}, \ldots$, but, of course, using the assumption $\mathbf{1\text{-}Con}_{T_0}$ which is stronger than all these statements. I conjecture that assuming a little more, namely Σ_1-RFN_T, we can extend this process to transfinite autonomous progressions (in the sense of [1]).

A possible source of misunderstanding may also be the fact that similar implications hold on various levels. For instance (3) is of this form. It enables us to iterate Rosser sentences, assuming the consistency of the initial theory. So the difference between the Rosser sentence, consistency (the Gödel sentence), ω-consistency and other possible variations is important, if we want to avoid false conclusions. This distinction is often disregarded in informal descriptions of Gödel's theorem .

Let us state and prove such implications for some principles that we have considered. The general form of these statements is

$$X_T \;\rightarrow\; X_{T+Y_T} \tag{6}$$

where X is the stronger and Y is the weaker principle. It is plausible that similar relations hold for other principles.

Proposition 3 *(1)* $Con_T \;\rightarrow\; Con_{T+Ro_T}$,

(2) Σ_{n+1}-$RFN_T \;\rightarrow\; \Sigma_{n+1}$-$RFN_{T+\Sigma_n\text{-}RFN_T}$,

(3) $\mathbf{1\text{-}Con}_T \;\vdash\; \mathbf{1\text{-}Con}_{T+Con_T}$,

(4) ω-$\mathbf{Con}_T \;\vdash\; \omega$-$\mathbf{Con}_{T+Con_T}$.

Proof. (1) is just a part of Rosser's theorem. (2) is proved in [13, Cor. 4.1.12].

To prove (3), first observe that $\mathbf{1\text{-}Con}_T$ implies Con_T. We shall use the fact that $\mathbf{1\text{-}Con}_T$ is equivalent to Σ_1-$\mathbf{Rfn}_T$. Assume Σ_1-$\mathbf{Rfn}_T$ and suppose $Pr_{T+Con_T}\ulcorner\varphi\urcorner$ for some φ in Σ_1. This means $Pr_T\ulcorner Con_T \rightarrow \varphi\urcorner$. The sentence inside is also Σ_1, hence we get $Con_T \rightarrow \varphi$. Now, using Con_T we get φ.

To prove (4) assume ω-$\mathbf{Con}_T$ and suppose $Pr_{T+Con_T}\ulcorner\exists x\varphi(x)\urcorner$. This means $Pr_T\ulcorner Con_T \rightarrow \exists x\varphi(x)\urcorner$. We shall rewrite this formula as

$$Pr_T\ulcorner\exists x(Prf_T(x,\ulcorner\bot\urcorner) \vee (Con_T \wedge \varphi(x)))\urcorner.$$

Applying ω-consistency of T to this formula we get

$$\exists x \neg Pr_T\ulcorner\neg Prf_T(\dot{x},\ulcorner\bot\urcorner) \wedge \neg(Con_T \wedge \varphi(\dot{x}))\urcorner. \tag{7}$$

Fix such an x. The consistency of T implies $\neg Prf_T(\underline{x}, \ulcorner\perp\urcorner)$. By Σ_1-completeness we get $Pr_T\ulcorner\neg Prf_T(\underline{x}, \ulcorner\perp\urcorner)\urcorner$. Thus (7) reduces to

$$\exists x \neg Pr_T\ulcorner\neg(Con_T \wedge \varphi(\dot{x}))\urcorner.$$

The formula inside is equivalent to $Con_T \rightarrow \neg\varphi(\dot{x})$, thus the whole formula is

$$\exists x \neg Pr_{T+Con_T}\ulcorner\neg\varphi(\dot{x})\urcorner.$$

This proves that ω-consistency holds for $T + Con_T$. $\square$

4 Conclusions

If we always justify a weaker principle by a stronger one, we are inevitably lead to so strong principles their truth is not evident to us. Consider for instance the hierarchy of the reflection principles Σ_n-RFN_T. The next step after all these principles is the uniform reflection principle for all arithmetical formulas Σ^1_0-RFN_T. In order to state this principle we need to be able to define the truth for all arithmetical formulas, which cannot be done in first order arithmetic. We need at least a fragment of the second order arithmetic. This is a big step. Natural numbers seem much more accessible to our intuition than subsets of natural numbers. The problem of the truth of the Continuum Hypothesis and several other problems about the continuum have not been resolved yet. Though these undecidable sentences are not directly linked with the reflection principle, it shows that we cannot be so confident anymore. As this is only a tiny part of the Zermelo-Fraenkel set theory, most people would go on, but at some stage everybody has to admit that the next principle is less likely to be true than the previous ones.

One of the arguments that Penrose uses is that whenever we accept T as our belief we accept also the soundness of T. He does not specify what exactly he means by the soundness. The only way to state it precisely that I see is to use some reflection principles. There is no theory which is closed off with respect to extensions by some reflection principle. As explained above we can expand T by taking stronger and stronger reflection principles, but at some point the principles will become too strong to be considered as obvious extensions of T. In the same way as people disagree on which set theory is safely consistent, people may disagree what extensions of an accepted theory T should be considered safe.

Thus it seems that our mathematical assumptions have a hierarchical structure like any other knowledge that we use. On the bottom there are statements that we believe are true without any doubts, while on the top there are doubtful statements which we have not been able to refute yet. We use less secure knowledge as a heuristic to guess the truth where we are not able to deduce it from more secure knowledge and data. (In our real life we mostly trust to our vision; when we cannot see the thing ourselves we may trust to somebody's report on it etc.) There are only a few, rather extravagant, logicians who doubt the consistency of Peano Arithmetic. On the opposite end the strongest assumptions are studied in set theory as *large cardinals*. The consistency of the largest ones is definitely more doubtful than the consistency of Peano Arithmetic, as at least in one case a proposed seemingly natural cardinal assumption had to be rejected as inconsistent.

The strong principles are rarely used directly, but we often use them unconsciously. Namely, the fact that no contradiction has been found for a strong principle strengthens our belief into a weak principle. For instance inaccessible, Mahlo and even measurable cardinals seem very safe, as no contradiction has been derived from considerably stronger principle in spite of extensive research. We also use stronger principles to produce "safe extensions" of weaker principles. We have demonstrated it on the example of the consistency principle and the reflection principle. The stronger one, the reflection principle, enables us to iterate extensions by the consistency. If we are too cautious, we do not have to accept a priori the reflection principle, but we may allow some consequences of it, namely, iterated consistencies. The feeling that we can always progress and make our assumptions stronger does not reflect our special ability, it is simply caused by the slow gradual decrease in our belief in their truth. Thus, after all, vagueness is present in our mathematical thinking, but not in the deduction process, it is in the decision which axioms we should accept.

References

[1] S. Feferman, *Transfinite recursive progressions of axiomatic theories*, J. Symbolic Logic 27 (1962), 259-316

[2] S. Feferman, *Theories of finite type related to mathematical practice*, in *Handbook of Mathematical Logic,* J. Barwise ed., North-Holland 1977, 913-971

[3] S. Feferman, *Gödel's program for new axioms: Why, where, how and what?* in *Gödel'96,* P. Hájek ed., Springer 1996, 3-22

[4] D. Hilbert, P. Bernays, *Grundlagen der Mathematik I,* Springer-Verlag, 1970

[5] D.G. Hofstadter, *Gödel, Escher, Bach: an eternal golden braid,* Basic Books 1979

[6] G. Kreisel, *Ordinal logics and the characterization of informal concepts of proof,* in *Proc. International Congress of Mathematicians,* J.A. Todd ed., Cambridge Univ. Press, 1960, 289-299

[7] G. Kreisel and A. Lévy, Reflection principles and their use for establishing the complexity of axiomatic systems, Zeitschr. f. Math. Logik 14, 1968, 97-142

[8] J.R. Lucas, *Minds, machines and Gödel,* Philosophy 36, 1961, 112-127

[9] R. Penrose, *The Emperor's New Mind, Concerning computers, minds, and the laws of physics*, Oxford Univ. Press, 1989

[10] R. Penrose, *Shadows of the Mind, A search for the missing science of consciousness,* Oxford Univ. Press, 1994

[11] R. Penrose, *The large, the small and the human mind,* Cambridge Univ. Press, 1997.

[12] H. Putnam, *a review of [10]* in Bull. AMS 32/3, 1995, 370-373

[13] C. Smorynski, *The incompleteness theorems,* in *Handbook of Mathematical Logic,* J. Barwise ed., North-Holland 1977, 821-865