STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 183–202
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.08
Article

ŠTĚPÁN HOLUB *

UNDERSTANDING, EXPRESSION AND UNWELCOME LOGIC

SUMMARY: In this paper I will attempt to explain why the controversy surrounding the alleged refutation of Mechanism by Gödel's theorem is continuing even after its unanimous refutation by logicians. I will argue that the philosophical point its proponents want to establish is a necessary gap between the intended meaning and its formulation. Such a gap is the main tenet of philosophical hermeneutics. While Gödel's theorem does not disprove Mechanism, it is nevertheless an important illustration of the hermeneutic principle. The ongoing misunderstanding is therefore based in a distinction between a metalogical illustration of a crucial feature of human understanding, and a logically precise, but wrong claim. The main reason for the confusion is the fact that in order to make the claim logically precise, it must be transformed in a way which destroys its informal value. Part of this transformation is a clear distinction between the Turing Machine as a mathematical object and a machine as a physical device.

KEYWORDS: mechanism, Gödel's theorem, Turing machine, hermeneutics.

The controversy surrounding the alleged refutation of Mechanism by Gödel's theorem is hard to approach. The discrepancy between the fact that the argument has been rigorously and unanimously rejected by logicians on one hand and the fact that proponents[1] are still defending it on the other hand, is striking. It indicates a deeper misunderstanding entrenched in the controversy. The verdict of

* Charles University, Faculty of Mathematics and Physics. E-mail: holub@karlin.mff.cuni.cz. ORCID: 0000-0002-6169-5139.

[1] By the term "proponents" (of the anti-Mechanist thesis) I will refer to thinkers who claim that Gödel's results refute Mechanism, mainly to John Lucas and Roger Penrose.

logicians was succinctly formulated by Hilary Putnam (1975a, p. 366): "misapplication of Gödel's theorem, pure and simple". The same critic later rejected a variant of the argument as a "sad episode in our current intellectual life" (Putnam, 1994). A more polite version of the same conclusion is the one by Stewart Shapiro (1998, p. 275): "My conclusion (perhaps slightly exaggerated) is that there is no plausible mechanist thesis on offer that is sufficiently precise to be undermined by the incompleteness theorems".

Nevertheless, the idea keeps provoking thinkers who again and again rush to add their take in the spirit of the opening sentence of John Lucas's original paper (1961, p. 112): "Gödel's Theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines". Lucas (2011) himself remained unimpressed by the criticism: "Since many critics are unaware of the argument, and are unlikely to look back at papers published some time ago, it is worth articulating the argument afresh". Lucas is willing to reiterate his feeling and he obviously believes that his extant answer to objections is sufficient. Apparently, something more is needed than Stewart Shapiro's (1998, p. 273) "modest aim of forging connections between different parts of [the] literature and clearing up some confusions, together with the less modest aim of not introducing any more confusions".

In this paper I will attempt to explain why proponents ignore logical arguments**,** and I will argue that they in fact want to establish a philosophical point which is not directly related to Mechanism. Instead they have in mind a fundamental feature of understanding which is the core tenet of philosophical hermeneutics, namely the insurmountable gap between any intended meaning and any of its formulations. The intention always contains more than what is captured by the expression. The proof of Gödel's theorem is particularly interesting specimen of this gap, in this case between our understanding of natural numbers, and its expression by a logical theory like Peano Arithmetic. However, this is a metalogical observation which does not disprove Mechanism, since it does not exclude the possibility that the human mind can be simulated by a formal system which absolutely surpasses human understanding**,** and therefore lies in a realm where no hermeneutics can be applied. Whether admitting such a formal system is reasonable or not becomes a futile argument without clear criteria. The main mistake of the proponents is that they appeal to a logical result which, as such, has no bearing on the dispute. They do not realize that in order to even make their claim intelligible**,** it is necessary to translate it into a logical language, whereby the proper hermeneutic insight is lost.

One of the first aspects of the translation is the definition of the Turing Machine which is equivalent to the recursive formal system. The difference between Turing Machines and their instantiations is stressed in the first section. The second section summarizes the discussion after the anti-Mechanist claim is properly formulated in logical terms and explains why its validity cannot be established. The third chapter develops the argument that the main feature of human mind that proponents want to highlight can be best described in terms of philosophical

hermeneutics. The chapter also explains why the hermeneutic feature cannot be suitably captured within the framework of logic. The fourth section analyzes the proof of Gödel's theorem in order to illustrate how an informal, that is, ordinary mathematical understanding of natural numbers is crucially responsible for its validity. This prompts considerations about the mutual dependence between formal and informal aspects of logic. The conclusion then points out the main lesson that can be learned from the curious discussion about Mechanism and Gödel's theorem, namely the need to respect methodical limitations of scientific disciplines.

1. Turing Machines and Robots

One of the standard reproaches against the anti-Mechanist thesis is that it is based on one of several problematic "idealizations" (Shapiro, 1998; Feferman, 2009). The main target is the "idealized human mind", whereas the idealization of a machine resulting in the concept of the Turing Machine is usually considered unproblematic, and the fact that the Turing Machine is not just a machine with infinite space and a flawless processor is easily underestimated.[2] However, only the idealized concept of the Turing Machine makes the proponents' argument possible because it corresponds to a recursive formal system in Gödel's proof. The "Turing Machine" is a mathematical entity in pretty much the same way as a natural number. The most important part of its definition is the transition relation which in turn is a finite set of quintuples, called instructions. The definition is then extended to a relation between "instantaneous descriptions", which, if the relation is deterministic, finally defines a (partial) map from natural numbers to themselves, called a recursively enumerable function. The way from "honest machines" to Turing Machines is therefore not a short and simple one. In the same way in which natural numbers are the mathematical conceptualization of discrete quantity, the Turing Machine is a mathematical conceptualization of a fully controlled process. We call such a process "mechanical" but that is only a metaphor.

It is worth noting that Turing (1937) originally used the word "computer" as a reference to a diligent and fully reliable clerk. Therefore, calling our electronic devices "computers" is just one, and an almost forgotten one, among anthropomorphic expressions (like "memory") we use for (electronic) devices. When we ask whether the human mind is a computer, it is therefore a kind of reversed metaphor, which actually asks whether the behavior of the mind can be exhaustively described as the activity of a diligent and reliable clerk. The precise sense of what "diligent and reliable" means is then mathematically captured by the notion of the Turing Machine. The nature of a "fully controlled process" is inde-

[2] Although already Shapiro correctly observed that the idealization performed in the definition of the Turing Machine is "similar to idealizations made throughout mathematics" (Shapiro, 2009, p. 275).

pendent of its realization, be it by the human mind or by a machine. The Turing Machine is therefore actually no machine in the strict sense. The frivolous use of metaphors concerning machines is an important contribution to the confusion surrounding these issues. When, for example, Paul Benaceraf (1967) calls the Turing Machine in question Maud, and notes that "she convinced herself […] of her own consistency" but that she "she shouldn't go around blabbing it", we may see it as a refreshing stylistic feature. However, the concept of the Turing Machine is so relatively recent (compared to, say, the concept of natural number), and so deeply connected to the idea of a "tape", and a "processing head", that it is difficult for many, at least for a majority of university students, to fully appreciate the fact that the Turing Machine is actually a mathematical object, even after they become familiar with other mathematical concepts and start to understand, for example, that working with a five-dimensional vector space does not require any magical sensory ability.

While physical computers people build are instantiations of intended Turing Machines (if nothing goes wrong), the opposite is much less obvious, and that is where metaphors may betray us. For example, I do not see any good reason for Krajewski's claim (2020, p. 35) that we would "have no doubt" that the robot Luke, a fictional result of a long robotic evolutional process, is a Turing Machine. Actually, as shown above, such a claim does not even make good sense. Neither is it clear to me how Luke's program (which is the Turing Machine we speak about) would be "investigated by human computer scientists" (p. 40). Two completely different problems are conflated here. The first one is how to obtain the "program" from a physical device (including a brain) at our disposal. That is, how to describe the behavior of the device by a finite set of states, and by a similarly finite set of transition rules governing their evolution conditioned by a finite set of possible instantaneous inputs. This problem, in addition to being close to hopeless, is not even remotely related to Gödel's theorem. Only then comes the additional question, namely whether we can understand what the Turing Machine obtained in the first phase "does", that is, to derive some relevant properties of the partial recursive function it defines. Even this is a daunting task, but at least it is somewhat related to Gödel's work.

If we call life-simulating artificial products "robots", we can then say that the problem is an inadvertent identification of robots with Turing Machines. While the question whether the Turing Machine can become conscious is as nonsensical as whether a sufficiently large natural number can, the question whether robots can eventually acquire mind is completely open, or at least it is a question which has hardly anything to do with formal logic.[3] Hilary Putnam dedicated several

[3] The anti-utopia drama *R.U.R.* in which Karel Čapek coined the word "robot" depicts the creation of robots as an invention on the chemical level. The robots are used as a labor force, and the question of computation is not particularly stressed. One of the "humanizing" aspects is the deliberate introduction of pain into their functioning, and the distinctive human feature, which robots eventually develop, is love, not understanding of Gödel sentences.

papers to the relation between minds, robots and Turing Machines, where he makes a similar point many times. Even in papers where he defended the thesis that "we are Turing Machines" he makes clear that the identity has to be understood as a "functional isomorphism" depending on a description of conscious life in terms of a finite set of discrete psychological states. What is at stake is a "functional organization", not "physical realization" (Putnam, 1975a, p. 373). Moreover, reflecting later on the implied condition of the existence of discrete states describing human experience, Putnam admitted—citing reasons one is tempted to describe as common sense—that his earlier "point of view was essentially wrong" (Putnam, 1975b, p. 298).

2. Why Proponents Are Wrong

Keeping in mind that we speak about recursive functions, not about robots, the intuitive appeal of the question is undoubtedly reduced for a non-mathematician or even a mathematician who is not a logician. Perhaps, its appeal should be reduced for the proponents themselves. In any case, the very meaning of the question now requires better clarification. What could it mean that minds can, or cannot, be explained as Turing Machines? The question must be reformulated in terms of the mind's output. Namely, the question becomes whether the set of all arithmetical propositions the human mind can in principle prove is or is not recursive. When pressed about the use of the incompleteness theorem in their claim, the proponents are therefore eventually forced to resort to a purely syntactic competition between the mind and the machine. The machine and the human mind will each produce sentences in a given formal system, and the human mind will always win by producing a true sentence (the Gödelian one) which the machine never will, unless the system is inconsistent. As Krajewski stresses (2020, p. 11), the content of the competition can be even reduced to establishing the solvability of Diophantine equations.[4] Since the precise conditions of this competition remain chronically unclear,[5] the focus turns to specifications of the idealized human mind. This necessarily leads to a construction of some abstract concept, ultimately mathematical but often accompanied by some playful theo-

[4] I found confusing, in this respect, the numerous remarks Krajewski makes about alleged circularity. For example, he says: "we should beware of a circularity: if we simply assume that the mind, which is self-conscious, does not operate according to […] rules, then we assume what we are supposed to prove by Lucas's argument, and the whole business with Gödel's theorem is superfluous" (2020, p. 19). Why so? Lucas's argument is that it can be shown beyond doubt from Gödel's theorem that the mind can outperform any Turing Machine in the field of solving Diophantine equations. This is independent of what we assume about the mind otherwise.

[5] Lucas's (2011) metaphor of the dispute against the mechanist in terms of the Oxford First and Second Public Examinations is just one example of how unclear it is.

logical terminology.[6] The discussion is already loaded by two dangerous ambiguities concerning machines (Turing Machines vs. robots) and the human mind (understanding vs. output).

Three basic technical facts govern the discussion. First, the most basic problem for the anti-Mechanist application of Gödel's theorem is the impossibility of proving the consistency of the considered system within the system itself (the impossibility is shown by the second incompleteness theorem). It is therefore not sufficient for proponents of human superiority to construct the Gödelian sentence, they first have to be able to show that the system is consistent, which is far from granted. This fact was quickly pointed out by many critics (it is also behind Putnam's "misapplication" remark), and it became one of the main points of contention. The second difficulty for the anti-Mechanist claim is that the construction of the independent Gödel sentence is itself algorithmic. That is, it can be performed by a suitable algorithm, although not the one corresponding to the examined theory. This leads to an infinite chase between Turing Machines, each new one "out-Gödeling" the previous one and being "out-Gödeled" by the next one. The third and technically most involved fact is a partial and final concession to proponents, called "Gödel's disjunction". It claims that either "mind is not a machine", that is, the set of sentences knowable by the "idealized human mind" is nonrecursive, or, if after all such a set is recursive, then the corresponding Turing Machine cannot be known, which in particular means that there are "absolutely unknowable" mathematical truths. This observation dates back to Gödel's own reflections on the matter which are often ridiculed for their perceived naïve "Platonism", but which nevertheless show both prudence and perspicacity concerning logical facts. Gödel's disjunction has proven to be a solid logical fact. Most important**,** it turns out that the second possibility, which represents a version of Mechanism, cannot be excluded by logical means. The technical layer of the literature on this provides a large variety of advanced and very interesting results in this direction, effectively warranting Shapiro's (cited above) "slightly exaggerated" informal conclusion. Moreover, the conclusion is shown not to be exaggerated at all in particular by the results presented in recent papers by Peter Ko-

[6] See the title of Benaceraf's paper (1967) or the skeptical remark of Peter Koellner (2018b, p. 476) about the "angelic mind". Shapiro (1998, p. 273) mocks this terminological manner when he writes: "A descriptive title for this paper would be 'Gödel, Lucas, Penrose, Turing, Feferman, Dummett, mechanism, optimism, reflection, and indefinite extensibility'. Adding 'God and the Devil' would probably be redundant". On the other hand, Peter Vopěnka entertained seriously the idea that the concept of god (or God) and his capabilities with respect to infinity helps to explain different conceptions of mathematics. The antic gods, corresponding to Christian angels, are able to see easily as small (or large) quantities as they wish, however always with the possibility to go deeper. The Christian God is on the contrary able to see the whole set of natural numbers or the absolute geometric point in one shot. This is obviously a variant of potential and actual infinity. However, Vopěnka both suggests that medieval theology directly influenced modern mathematics, and tries to use the theological explanation as a common sense basis for the non-standard analysis and for its practical use (cf., for example, Vopěnka, 2010).

ellner (2018a; 2018b). Both the strength and the weakness of such technical results is that they are, by definition, results about some formalized versions of the Mechanist claim. Although the details are sophisticated, the nature of the results is fairly straightforward. Since provability has its precise technical meaning, it remains to identify formal counterparts for truth and knowability. This requires the introduction of predicates or operators T and K, to formulate suitable axioms for them, and then to show, by standard (or rather advanced) logical means corresponding facts, namely the relative (in)consistency of certain scenarios. As indicated, all these results are devastating for the proponents in the sense that carefully formulated versions of Mechanism informed by Gödel's disjunction are logically consistent (provided Peano Arithmetic is) in all situations one can think of.

To provide those logical achievements here in more detail is both unnecessary and insufficient for the simple reason that the proponents themselves seem to ignore them by plainly dismissing the whole glorious technicality in favor of alleged informal evidence against the second option of Gödel's disjunction. The discussion could be closed here and shifted to a different kind of philosophical investigation of the mind. The trouble is that the proponents want to base their philosophical argument on, of all things, Gödel's theorem. They insist that their original insight, if properly understood, is valid despite the objections.[7]

We may try to summarize the whole controversy as follows. Proponents assume (implicitly and sometimes explicitly) a self-evident capacity of the human mind ("getting hold", "twigging", "truth-divining", see note 7), which can briefly be called u n d e r s t a n d i n g. They further see the ability to understand as an obviously non-mechanical attribute. This is essentially what Descartes tried to say in his oft-quoted anti-Mechanist argument,[8] or what John Searle illustrates by his Chinese room argument. Descartes apparently considered the test of the presence of understanding to be a matter of course, which is not the case anymore for us who know modern computers. In any case, we simply know (or feel)

[7] Relevant quotes are, for example:

"There is a way of arguing that commends itself to those possessed of minds, who get the hang of the Gödelian argument, and twig that they can apply it, suitably adapted, in each and every case that crops up. Mechanists may refuse to see the general case, and, acknowledging only knock-down arguments, will have to be knocked down each time they put forward a detailed case: minds can generalise, and will realise that defeat for the Mechanists is always inevitable" (Lucas, 2011).

"As to the very dogmatic Gödel-immune formalist who claims not even to recognize that there *is* such a thing as mathematical truth, I shall simply ignore him, since he apparently does not possess the truth-divining quality that the discussion is all about" (Penrose, 1999, p. 582).

[8] "For it is highly deserving to remark, that there are no men so dull and stupid, not even idiots, as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and that on the other hand, there is no other animal, however perfect or happily circumstanced, which can do the like" (Descartes, 1637, Part V).

we are conscious and express meanings, and there is no real argument about this. What becomes unclear is whether the existence of understanding can be conclusively proven exclusively on the syntactic level, that is, on the level of produced signs. This yields the question whether syntactic rules can simulate understanding successfully, that is, whether the same syntactic output can be obtained without the corresponding understanding. Here an apparently equally obvious proposition arises, namely that Gödel's incompleteness theorem proves conclusively the impossibility of such a simulation. This is the core anti-Mechanist thesis. The latter proposition is nevertheless wrong, since the second possibility in Gödel's disjunction remains unrefuted: it may be the case that the entire output corresponding to the (human) understanding is successfully simulated by purely syntactic rules, namely rules that (forever) transcend the (human) understanding in question. This is the state of the art from the technical point of view, which, however, makes nobody happy. Opponents cannot concede the thesis while proponents understandably feel that the objection misses the point. Lucas's objection could be formulated as follows: The syntactic rules mentioned above must make some sense, namely as rules. It is irrelevant, Lucas can insist, whether the human mind can or cannot understand them, in any case they are understandable "in principle", understandability is part of their being rules. Consequently, in order to save the point, proponents are forced to adopt some kind of metaphysical commitment concerning formal systems and the capacity of human mathematical understanding, which, however, have no clear backing in Gödel's theorem.

3. What Proponents Want to Say

The proponents were lured into an incorrect logical claim by the necessity to formulate their claim as a thesis that permits logical proof, which in turn implied dubious metaphysical assumptions. I want to suggest that the real point the proponents are after is something different, namely the inexhaustibility of understanding by expression, of meaning by syntax. Let me start by illustrating the uncertain relation between understanding and Turing Machines (or formal systems) first with an example of a finite structure like chess, and then with the question of consistency.

From the point of view of the present anti-Mechanist argument, chess is a trivial case of a finite directed graph of legal positions with edges representing moves. Every possible claim about chess is trivially decidable by an exhaustive search. On the other hand, it is safe to say that as long as human competitive chess will exist, we shall continue to speak about the understanding of a position in chess. In order to assess how such an understanding relates to computations done by a Turing Machine, let us compare a brute force algorithm with the sophisticated engines we can use today that define an evaluation function and optimize it within a large, but still limited search space. The evaluation function incorporates a formalization of the understanding of chess by top players, or, it is blindly inferred from a huge number of matches (in cases such as AlphaZero

deep learning program). The evaluation function is the closest parallel to a "computer's understanding" of the game, and it is what practical artificial intelligence is all about. Nevertheless, if we ignore questions of computational complexity (which have no significant place in the anti-Mechanist controversy), the tricky nature of evaluation function becomes irrelevant. The brute force algorithm, which contains no advanced intelligence and could be written by any decent undergraduate student, becomes unbeatable. This is a rather trivial illustration of the fact, that by "understanding" we mean something else then blind syntactic ability. If we want to interpret "artificial intelligence" as an "understanding" possessed by Turing Machines, we either have to consider computational complexity, or to explain why understanding of finite (albeit very large) structures is substantially different from understanding of infinite ones.

A touchstone for what the role of understanding is within formal logic is the question of consistency. Aristotle, in his original formulation of the principle of non-contradiction, argued that contradiction must be excluded because it destroys meaning.[9] It is completely unclear what somebody says, or whether he says anything at all, if the same claim is asserted and denied in the same time and the same sense. The care with which the sameness of the two claims is stressed underlines how we usually deal with an inconsistency. We either try to repair it on the formal level of expression (as a typo), or, when the misprint is excluded, we try to search for a deeper distinction which would make the apparent sheer contradiction comprehensible. This is more than "overcoming the contradictions by pointing to the metaphorical character of expressions" as Krajewski suggests in one place (2020, p. 22); unless "metaphor" is understood not as a "mere metaphor" but as a substantial feature of any meaningful speech. Let us consider the seminal example of a set of all sets that are not elements of themselves. There may be an argument about whether the very expression "being its own element" makes sense. The answer will depend on what exactly we mean by "incidence", that is, by "being an element of". We may try to capture the exact meaning by various ways of reflection, for example by some kind of Husserlian "eidetic variation". Formal logic proposes to investigate the question on the syntactic level of propositions that include the word "incidence". We are invited to pretend that we have no idea at all what the sign $\in$ means. We just understand how it can be manipulated (note for further purposes that even this is a kind of understanding). Eventually, we discover a sentence that can be derived, according to the rules, as well as its negation. What shall we do? Formally speaking, we just discard the theory. On practical level, some kind of "correction" takes place, so often invoked in the anti-Mechanist controversy. The set in question certainly

[9] "If on the other hand it be said that 'man' has an infinite number of meanings, obviously there can be no discourse; for not to have one meaning is to have no meaning, and if words have no meaning there is an end of discourse with others, and even, strictly speaking, with oneself; because it is impossible to think of anything if we do not think of one thing…" (*Metaphysics*, IV, 1006b). See my paper (Holub, 2004) in Czech for a discussion of Aristotle's approach.

cannot in the same time and in the same sense be and not be its own element, independently of what incidence means. That makes no sense. However, it does not imply that a set cannot be its own element. Although making the formula "$x \in x$" itself contradictory is one possible (and standard) solution, it is an overcautious one. There are theories which allow sets to be elements of themselves. The collection of sets that are not elements of themselves is then just not itself a set. If we consider the last conclusion paradoxical, then we have not taken the formalization seriously. During the formalization, we were asked to forget completely that variables are supposed to refer to "collections". In fact, there is a standard technical (meta)term for this kind of collection, namely a proper class. If this is not a proof of a specific mathematical sense of humor, it is at least a proof of a pragmatic approach which cares as little as possible about formal contradictions, and instead is driven by understanding.

In contrast to the essentially infinite nature of both Turing Machine computation and the consistency requirement, the likely original motivation of the proponents is relevant already for human understanding in its finite form. We have to abandon the misleading and unclear idea of simulation and focus instead on the tension between expression and its meaning. This happens to be the starting point of an area of philosophy as alien to formal logic as *philosophical hermeneutics*. According to its main exponents, the core tenet of philosophical hermeneutics is *verbum interius*,[10] or the *surplus of meaning*,[11] the fact that the meaning intended by the speaker or writer never perfectly matches the linguistic expression. This precludes a direct approach to the intended meaning for an interlocutor or reader, making an interpretation necessary. Moreover, there is no pure "original intention", independent of the expression, for the speaker either. In order to fix any meaning, it is necessary to express it. The need for interpretation therefore applies to all thinking which becomes, in Plato's words, an inner dialogue of the mind. The dialectics of understanding and expression is grounded in the unique perspective of the author and the unique context of the locution as opposed to the stability of the expression, which allows others to share the intended meaning, as well as the authors to return, possibly with a surprise, to their own previous thoughts. The gap between expression and meaning is revealed by a reflection on the expression, and the comparison with meaning it allows. The expression is a transparent medium leading directly to the meaning in the case of a successful understanding. We become aware of the expression when the understanding is disrupted. The expression then loses its transparency, becomes visible as an independent reality, and an interpretation is needed in order to reestablish understanding. Such an interpretation adds new expressions that may help to elucidate the original meaning but, at the same time, they themselves may be-

[10] See the foreword to Grondin's (2011), where the author quotes his discussion with H. G. Gadamer.

[11] See, for example, Ricoeur's (1976).

come unclear. The process then continues, until an understanding needed for practical needs of the particular situation is achieved.

Hermeneutics shares with logicism the suspicion concerning a direct approach of consciousness to itself, in some kind of transcendental reflection not mediated through any expression. Paul Ricoeur replaces the self-transparent Cartesian *cogito* with a *cogito brisé*, a broken consciousness. Formal logic is a deliberate strategy to make the expression fully "opaque", fully devoid of meaning. Its full focus is on the syntactic rules. We remarked above that even in this case an understanding of the formal system *qua* formal system is required (for example, understanding of how well-formed formulas can be obtained). The formal system thus becomes a mathematical object in an ordinary sense, which substitutes for the original one (like natural numbers) and which can be informally, or again formally, investigated. However, the investigation should receive no guidance from the original, motivating understanding, lest be misled by it. It was Frege's and Hilbert's hope that restricting understanding in this radical way will eventually yield a better grasp of the original meaning. The hope is that syntactic or logical rules, while being simpler to control, will nevertheless fully substitute for the suspended meaning. This hope was frustrated by Gödel. Even in mathematics, the understanding always means more than what its formulation says explicitly.

4. Informal Mathematics in Gödel's Proof

The proof of Gödel's theorem reveals very clearly the above described hermeneutic principles through the relation between formal expressions of Peano Arithmetic on one side, and the informal mathematical understanding on the other side. I will show this by an analysis of the technical content of the proof. The main goal of this analysis is to trace informal mathematical aspects of the proof. "Informal mathematics" should be understood as ordinary mathematics, which in fact uses formalism quite heavily, but which is nevertheless not formal in the logical sense. In other words, "informal" means mathematical but at the same time meta-logical.[12]

Gödel's incompleteness theorem claims that any recursive formal theory that is sufficiently strong contains a sentence such that neither the sentence nor its negation is provable in the system. The theory in question can be some theory designed to capture our understanding of natural numbers, for example Peano Arithmetic.

The proof of the theorem is based on three technical ingredients.[13] The first one is the famous Gödel numbering which establishes a correspondence (possi-

[12] I think this is what Gödel (1931, p. 176) has in mind when he speaks about "metamathematische Überlegungen".

[13] An excellent self-contained exposition of the main structure of the argument, spanning only five pages, can be found in the first chapter of Smullyan (1992, pp. 5–9).

bly one-to one) between formal expressions in the language of the theory, and natural numbers. This is a crucial step since in this way formulas are transposed from the level of language to the level of objects the language is supposed to describe. This allows us to eventually interpret the theory in a way that yields some information about the theory itself. It should be stressed, however, that the correspondence between numbers and formulas is realized on the meta-level, by the mathematician writing the proof. Moreover, the existence or meaningfulness of the very structure of natural numbers that we use to encode formulas is of course in no way guaranteed *a priori* by the theory that is designed to describe them. We rely on our pre-formal (in the logical sense) meta-understanding.

The second main ingredient of Gödel's theorem is the diagonalization, which is also the core of other related cornerstones of modern mathematics, such as the existence of algorithmically undecidable problems**,** and the concept of higher infinities beyond countable infinity. The simplicity of the idea deserves to be stressed and kept in mind. In fact, it is recommended to contemplate the basic nature of the diagonal argument as an antidote whenever one is tempted by the "mystical charm" (Krajewski, 2020, p. 15) of Gödel's theorem or related results. The finite version of the diagonal argument provides an elegant constructive form of the fact that the number of sequences of length n is larger than n (provided there are at least two distinct symbols). It is always straightforward to exhibit a particular missing sequence, namely the "negated diagonal", that is, the sequence whose i-th element is a symbol distinct from the i-th element of the i-th sequence of the list. This idea can be extended to any countably infinite list of infinite sequences, yielding Cantor's proof for the uncountability of the continuum.[14]

The third main ingredient of Gödel's theorem is expressibility, the ability to describe certain important features of the language in terms of the language itself. Here we essentially use the above introduced encoding. More careful formulation should therefore be that the encoding is extended to more complex linguistic expressions (ultimately to formal proofs), and the numbers that represent expressions with desired properties are captured by suitable formal expressions. Specifically, the theory must be able to formulate "n-th expression to which number n is substituted" (realizing the diagonal idea), and, most prominently, "m is a proof of n-th expression". Once more, m actually is a number, which encodes a formal proof of the n-the expression. Construction of formulas representing the above properties is the technically difficult part of the proof,

[14] Let me remark that Cantor's theorem can serve either as a starting point for doubts about actual (as opposed to potential) infinity, or as the entrance gate to "Cantor's paradise" of Set Theory. Set Theory then postpones its moment of reflection to the problem of the "universal diagonal" of the collection of sets that are not elements of themselves, which is famously contradictory if accepted naively. Set Theory could thus be characterized as the grey area created by the diagonal argument and extending between full acceptance of actual infinity and the rejection of contradiction.

requiring a logician of Gödel's greatness.[15] Again, it must be stressed that by proof we mean a formal proof here, that is, a sequence of formulas obtained by successive elementary derivation steps. This observation can hardly be overestimated. In fact, the difference between formal and informal proof is probably the most contentious aspect of the debate. Just as the Turing Machine is no machine although it can be instantiated by one, formal proof does not prove anything although it is designed to reflect a full-blooded proof and can be interpreted as such.

What exactly is required from the formal theory to be able to express notions needed for the incompleteness theorem can be investigated in several ways. The standard mathematical way a theory is shown to be too weak, and therefore complete and decidable, is q u a n t i f i e r e l i m i n a t i o n. This is related to the fact, we pointed out above, that finite structures are decidable trivially. Undecidability always arises due to the presence of quantified formulas, formulas which dare to claim something about all objects. Quantifier elimination reveals the weakness of a theory by showing that each such formula is in fact equivalent to one without quantifiers. The theory is shown to be too weak to be able to say something universal.

Seen from this perspective, the gap between the decidable Presburger Arithmetic and the undecidable Peano Arithmetic becomes curious. The difference between them is the presence of multiplication. It turns out that speaking about natural numbers in terms of addition only does not allow anything to be said universally. Krajewski (2012) considers the appearance of undecidability when multiplication is added to be one example of "emergence" in mathematics, as a fact that remains irreducibly surprising even for an expert. Let me attempt a speculative explanation of this phenomenon.[16] It may be argued that multiplication is the place where natural numbers start to apply to themselves. While the basic role of natural numbers is to count objects (be they apples or abstract units), in multiplication the counted objects become numbers themselves. No more five times an apple, instead five times four.[17] The four is suddenly not only a quantitative property of a particular assemblage, it becomes a proper object which itself deserves to be counted. This can be therefore seen as the required threshold of "self-reflection".[18]

Using the above three ingredients, Gödel is able to find an independent sentence, a sentence which can neither be proved nor disproved in the formal system. Formally speaking, we have a pair of formulas (the sentence and its negation),

---

[15] There is, of course, the difference between the difficulty of a proof, and the difficulty of discovering it. In today's form the full proof can be included in an undergraduate course. Gödel himself (1931, p. 173) calls the independent sentences he derives in his famous paper: "Relativ einfache Probleme aus der Theorie der gewöhnlichen ganzen Zahlen".

[16] I am indebted to a remark by Kateřina Trlifajová for this idea.

[17] Of course, both the multiplicand and multiplier must be allowed to be universally quantified. Multiplication by a given individual number can be expressed as a sum.

[18] I wonder whether this speculation can be somehow supported by the analysis of quantifier elimination.

such that both of them are unprovable, that is, they cannot be obtained from other specific formulas (called axioms) in a prescribed way. This purely formal fact does not sound very interesting. Its real importance depends on the interpretation we give to the formulas. More specifically, we interpret sentences as claims about natural numbers that can be true (or false). Also, we interpret formal proof as an object that faithfully captures ordinary mathematical reasoning, which, in turn, depends on the truth preserving quality of certain reductions.

Finally, we are convinced that the theory in question (Peano Arithmetic to start with) is consistent.[19] This is a particular case of the way mathematicians tend to deal with possible inconsistency, which we have discussed above. The consistency of Peano Arithmetic is a particularly bold assumption, and some serious mathematicians have even sincerely doubted that it is the case.[20] The point is that Peano Arithmetic contains infinitely many axioms within the scheme of induction. In particular, it contains the induction claim for arbitrarily large and complex formulas, even for those we shall never be even able to read, let alone to understand what they say.[21] However, even if it turned out that Peano Arithmetic is contradictory, it would not, as Krajewski correctly observes (2020, p. 22), necessarily disturb ordinary mathematics in any significant way. We can imagine that an extremely huge proof of contradiction would somehow miraculously appear somewhere on the internet. It would be fairly easy to verify, using computers, that the contradiction is genuine. Nevertheless, it would involve a lot of extremely complicated instances of the scheme of induction. Lucas believes that we would eventually be able to sort things out, an example of the depth of his optimism. From the practical point of view, however, it would just mean that some of the involved axioms should be prohibited. Undoubtedly, a new field of research would be created to investigate which one, but ordinary mathematicians would be, at best, just more conscious of what kind of induction they use.

Nevertheless, let us believe with Lucas that "we" (whoever that is) are "in principle" (whatever that means) able to understand and even to verify the validity of all axioms. Combining this with the truth preserving quality of formally logical inferences, we are ready to claim that no arbitrarily large derivation within the whole monstrous theory can lead to a contradiction simply because such a contradiction could, "in principle", be translated into firm evidence of the fact that zero equals one. Finally, granted all this, we are able to contemplate the insufficiency of our theory (Peano Arithmetic) to exhaust our intuition.

---

[19] More precisely, we believe that it is ω-consistent, which excludes the possibility that a provable universally quantified formula is at the same time disprovable for all numerals.

[20] See, for example, (Nelson, 2006). The intransigence of Nelson's views has certainly been compromised by his mistaken announcement that he actually proved the inconsistency of Peano Arithmetic.

[21] Nelson (2006) points out further difficulties related already to induction on relatively simple formulas, which illustrate that our belief in natural numbers is far from self-evident.

It is not that difficult to understand why this magnificent proof leads Lucas to celebrate "getting the hang of the argument" and "twigging that we can apply it" or Penrose to brag about a "truth-divining quality" (see remark 7) that are supposed to establish the superiority of creativity over rule-following.[22] But even if we enjoy joining Lucas in his exaltation of the scintillating human mind, we have to return to our question about the exact contribution of Gödel's theorem here. We have to do so because we have seen how rather than support optimism, the validity of the theorem turns out to depend on it. At best, its proof is one among many occasions to experience mathematical understanding at work. It also makes clear that our understanding is not exhausted by theorems provable in Peano Arithmetic, or in any theory for which we are able to perform a similar construction, provided that the theory is consistent. But it does not exclude the possibility that our mathematical understanding is governed by a formal system for which we are unable to carry out the proof, since we are unable to understand it. We have seen that the main difference between proponents and critics is whether they care about the latter qualifications. While critics expected that the argument will deal with them, proponents instead took them for granted. But then Gödel's theorem is just an instance of mathematical understanding, which can be philosophically investigated but whose specific content provides no particular philosophical contribution. The argument is flat, and it is understandable that Lucas wants to turn the page.[23] Frege's original objective was to explain the conceptualization involved in mathematics by means of logic. The fact that in Gödel's theorem logicism defeats itself, as it were, by its own means is undoubtedly an epochal result. On the other hand, the discussion about the central objective of logicism has gone on since the sixties within neo-Fregeanism and neologicism.[24] Lucas's paper could have been part of that discussion, less famous but philosophically more substantial, had it started with "Gödel's Theorem seems to me to prove that natural numbers cannot be described purely logically…"

The failure of the project induced by Gödel's theorem represents a vindication of (Kantian) intuition. However, as soon as formal logic becomes an established mathematical discipline, it can be cultivated independently of its philosophical foundations, as can any other mathematical discipline, like arithmetic or geometry. Gödel's theorem may then lead unprepared students astray to nonsensical conclusions of the following kind. The failure of the attempt to capture fully the whole formal truth about natural numbers, seen from the point of view of formal logic, casts doubts not primarily on the logic itself but rather on our original intuition about natural numbers. Is there something like the standard model at

[22] See: "[A]lthough Gödel cannot make us scintillate, he does show that scintillation is conceptually possible. He shows us that to be reasonable is not necessarily to be rule-governed, and that actions not governed by rules are not necessarily random" (Lucas, 2011).

[23] For the same reason, it is unclear why Krajewski's "Theorem of Inconsistency of Lucas" should be described as "unexpected" (2020, p. 32).

[24] See for example (Kolman, 2005) for a polemic against the neologicism.

all? Is the standard model somehow distinguished among other models? Is it possible to explain the standard model in set theoretical terms? Which set theoretic model is appropriate and why? A working logician sometimes seems to feel much more comfortable when speaking about nonstandard models, and some even believe that Gödel's incompleteness theorem actually shows that there is nothing like the standard natural numbers.[25] However, the proof of the theorem, as we stressed, is based on the interpretation of formal sentences as claims about natural numbers. What natural numbers? If the logical analysis deconstructs our concept of natural numbers, then the deconstruction itself is undermined as far as it depends on the interpretation of natural numbers. Where do we stand then? The space for sophistry and wild mystical comments is open in this aporia (and the opportunity is amply seized). For example, it is a basic "ordinary mathematical" conclusion that the Gödelian formula, which declares itself to be unprovable, indeed is unprovable, and therefore true. If it were provable, then it would be true, and therefore unprovable (since that is what it says under the interpretation), which is an (informal, ordinary mathematical) contradiction.[26] Does this argument work anymore when we are not sure about the standard interpretation? Since the formula is independent of the axioms of the theory under investigation, there is a (nonstandard) model in which it is provable. Does Model Theory magically allow (formal or even informal) contradiction?

These are questions that require a calm reflection on the technical ingredients of Gödel's theorem listed above, combined with keeping in mind that, when proving Gödel's theorem, we are doing "ordinary mathematics" with ordinary natural numbers. Formulas are just sequences of symbols with no magic power to destroy our mathematical understanding. These formulas are actually objects of our mathematical understanding as much as natural numbers, which is particularly apparent in the encoding of formulas by numbers. Our knowledge, then, is fully dependent on the interpretation we give to formulas as claims about natural numbers, and on the truth preserving quality of derivation. The sentence is true, in natural numbers, because otherwise we would obtain an inconsistency in our understanding of natural numbers and of truth derivation. The alleged proof of the sentence would be inconsistent with the claim the sentence makes about its own unprovability. The latter, recall, depends on the encoding. The sentence claims that there is no number with certain subtle properties expressed by a complicated formula. However, the alleged proof of the sentence, if encoded, will yield a number with exactly those properties. In order to see this, we have to "get

[25] Concerning set theoretic models let me mention a paper by Paul Benacerraf (1965, p. 73) which happens to conclude: "They think that numbers are really sets of sets while, if the truth be known, there are no such things as numbers; which is not to say that there are not at least two prime numbers between 15 and 20".

[26] See already Gödel's original paper: "Aus der Bemerkung, dass [R(q); q] seine eigene Unbeweisbarkeit behauptet, folgt sofort, dass [R(q); q] richtig ist, denn [R(q); q] ist ja unbeweisbar (weil unentscheidbar). Der im System PM unentscheidbare Satz wurde also durch metamathematische Überlegungen doch entschieden" (1931, p. 176).

the hang" of the proof, in Lucas' words. Finally, the Gödelian sentence together with all provable sentences of the original theory form a consistent system, which therefore has a model. This model is a mathematical structure, which looks much like natural numbers, but it contains an element which represents the proof of the Gödelian formula in the original theory. Nevertheless, this brings about no inconsistency, since the element representing the (non-existing) proof of the Gödelian formula is not a standard number, therefore it yields no sequence of formulas we would accept as a proof in "ordinary mathematics". The firm basis of all this in informal mathematics is obvious.

5. Conclusion

The troubled dispute about Mechanism inspired by Gödel's theorem is an instructive example of difficulties resulting from the lack of respect for methodical limitations of different scientific areas. Three levels are at play: mathematical logic (formal arithmetic), (informal) mathematics of natural numbers, and the philosophical reflection on both these disciplines. Since the anti-Mechanist claim has a philosophical nature, after its forced transition to the field of formal logic it suffers from not sufficiently distinguishing between technical results on the one hand, and the philosophical reflection on their significance on the other. While moving between "levels" and "meta-levels" is well established within formal logic, it typically happens within the discipline itself. Model Theory builds models of one theory using another one, obtaining thereby essentially relative results. Asking philosophical questions transcending the discipline as a whole is often seen with suspicion by logicians, if not directly dismissed as a delusion.[27] Frege (1998, p. XII) was well-aware of this predicament when he was skeptical about prospects of his own work.[28] However, the anti-Mechanist thesis is precisely a reflection on philosophical consequences of purely technical results. A seemingly paradoxical principle applies to such attempts. In order to estimate the relevance of philosophical, informal conclusions drawn from a technical theorem,

[27] Peter Koellner (2018b, p. 476), for example, dismisses concepts of "absolute provability" and "knowability by the idealized human mind" as "not sharp enough for our questions […] to have definite sense and determinate truth-values", but nevertheless continues: "With the above discussion in place, I would like to once again suspend the above skeptical considerations and assume, for the sake of argument, that the concepts of 'absolute provability' and 'knowability by the idealized human mind' are definite".

[28] "Sonst sind die Aussichten meines Buches freilich gering. Jedenfalls müssen alle Mathematiker aufgegeben werden, die beim Aufstossen von logischen Ausdrücken, wie "Begriff", "Beziehung", "Urtheil" denken: *metaphysica sunt, non leguntur!* und ebenso die Philosophen, die beim Anblicke einer Formel ausrufen: *mathematica sunt, non leguntur!* und sehr wenige mögen das nicht sein. Vielleicht ist die Zahl der Mathematiker überhaupt nicht gross, die sich um die Grundlegung ihrer Wissenschaft bemühen, und auch diese scheinen oft grosse Eile zu haben, bis sie die Anfangsgründe hinter sich haben. Und ich wage kaum zu hoffen, dass meine Gründe für die peinliche Strenge und damit verbundene Breite viele von ihnen überzeugen werden".

it is absolutely necessary to investigate the nature of the technicality involved. In other words, the conceptualization involved in the formal logic has to be prominently taken into account. Attempts to draw informal conclusions from a formal argument understood informally, are, in my opinion, the essence of the imprecision in thinking that causes a big part of the sadness of the corresponding episodes of our intellectual life.

The tension between formality and informality creates a difficulty which has its impact already on the superficial level of the external organization of relevant papers. They typically contain long sections of technical explanations using formal language.[29] The technical achievements then have to be interpreted, translated into ordinary language and their relevance has to be established. The technical language also typically contains lots of terms with suggestive non-technical meanings, terms that tend to permeate the informal discussion although the non-technical meaning may very poorly reflect the term's technical function. Real numbers, to take a trivial example, are no more real than natural numbers. While nobody is likely to draw philosophical consequences from the term "real number", the term "provable" in our context turned out to be significantly less safe. We have discussed the example of the word "machine".

I have attempted to show that in the Mechanist controversy the ordinary human mind, and its capacity to understand, is the first casualty of the battle which is officially waged for its sake. It simply turns out that mathematics, or at least formal logic, has no good tools to capture the cherished superiority of understanding.[30] The anti-Mechanist argument is a misguided effort to vindicate the capability of a particular human mind by means of the idealized one. We have seen how heavily the proof of Gödel's theorem depends on the informal understanding of arithmetic. The proponents make a quixotic attempt to translate the evidence for the superiority of the informal understanding to the formal level. They want to use the failure of logicism to prove at least this failure in the logically water-proof way. Gödel's incompleteness theorems are a kind of touchstone for the ambition of formal logic to substitute syntax for semantics for good. Lucas and Penrose are tragic heroes of this fight. They are driven by the obvious superiority of meaning over the syntax. However, they make a foolish choice of attempting to prove this superiority by the very means of syntax. While the Mechanist wants to reduce meaning to the pure manipulation of symbols, Lucas and Penrose want to vindicate the superiority of meaning—by pure manipulation of symbols. We are however well advised to spend the finite resources of our creative mind on something more reasonable than syntactic competitions with Turing Machines.

---

[29] The remark by Paul Benaceraf (1967, p. 13) often applies: "I trust that the following exposition will prove too elementary to be of any interest to those who are familiar with the logical facts, and too compressed for those who are not. For the sake of future reference, however, it must be done".

[30] Cf. (Feferman, 2009, p. 213): "[I]t is hubris to think that by mathematics alone we can determine what the human mind can or cannot do in general".

REFERENCES

Benacerraf, P. (1967). God, the Devil, and Gödel. *Monist*, *51*(1), 9–32.

Benaceraf, P. (1965). What Numbers Could Not Be. *Philosophical Review*, *74*(1), 47–73.

Descartes, R. (1637). *Discourse on the Method*. Leiden. Retrieved from: http://www.gutenberg.org/files/59/59-h/59-h.htm

Feferman, S. (2009). Gödel, Nagel, Minds, and Machines. *Journal of Philosophy, 106*(4), 201–219.

Frege, G. (1998). *Grundgesetze der Arithmetik*. Olms Verlag.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, *38*, 173–198.

Grondin, J. (2011). *Einführung in die philosophische Hermeneutik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Holub, Š. (2004). Aristotelův princip sporu ve čtvrté knize Metafyziky. *Reflexe*, *25*, pp. 71–80.

Klein, J. (1969). *Greek Mathematical Thought and the Origin of Algebra.* Cambridge, Mass.: M.I.T. Press.

Koellner, P. (2018a). On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose. *Journal of Philosophy*, *115*(7), 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized, II: Penrose's New Argument. *Journal of Philosophy*, *115*(9), 453–484.

Kolman V. (2005). Lässt sich der Logicismus retten. *Allgemeine Zeitschrift fur Philosophie*, *30*(2), 159–174

Krajewski, S. (2020). On the Anti-Mechanist Arguments Based on Gödel's Theorem. *Studia Semiotyczne*, *34*(1), 9–56.

Krajewski, S. (2012). Emergence in Mathematics? *Studies in Logic, Grammar and Rhetoric*, *27*, 95–105.

Lucas, J. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112–127.

Lucas, J. (2011). *The Gödelian Argument: Turn Over the Page*. Retrieved from: http://users.ox.ac.uk/~jrlucas/Godel/turn.html

Nelson E. (2006) *Warning Signs of a Possible Collapse of Contemporary Mathematics*. Retrieved from: https://web.math.princeton.edu/~nelson/papers/warn.pdf

Penrose, R. (1999). *The Emperor's New Mind*. Oxford: Oxford University Press.

Putnam, H. (1975a). Minds and Machines. In: H. Putnam (Ed.), *Mind, Language and Reality: Philosophical Papers* (vol. 2, pp. 362–385). Cambridge: Cambridge University Press.

Putnam, H. (1975b). Philosophy and Our Mental Life. In: H. Putnam (Ed.), *Mind, Language and Reality: Philosophical Papers* (vol. 2, pp. 291–303). Cambridge: Cambridge University Press.

Putnam, H. (1994). The Best of All Possible Brains? Review of Roger Penrose, *Shadows of the Mind*. *New York Times Book Review*, *144*, 7.

Ricoeur, P. (1976). *Interpretation Theory: Discourse and the Surplus of Meaning*. Fort Worth, Texas: Texas Christian University Press.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *The Bulletin of Symbolic Logic*, *4*(3), 273–302.
Smullyan, R. M. (1992). *Gödel's Incompleteness Theorems*. Oxford: Oxford University Press.
Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265.
Vopěnka, P. (2010), *Calculus infinitesimalis*, *pars prima*. Praha: OPS.