STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 243–266
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.11
Article

PAULA QUINON [*]

THE ANTI-MECHANIST ARGUMENT BASED ON GÖDEL'S INCOMPLETENESS THEOREMS, INDESCRIBABILITY OF THE CONCEPT OF NATURAL NUMBER AND DEVIANT ENCODINGS

SUMMARY: This paper reassesses the criticism of the Lucas-Penrose anti-mechanist argument, based on Gödel's incompleteness theorems, as formulated by Krajewski (2020): this argument only works with the additional extra-formal assumption that "the human mind is consistent". Krajewski argues that this assumption cannot be formalized, and therefore that the anti-mechanist argument – which requires the formalization of the whole reasoning process – fails to establish that the human mind is not mechanistic. A similar situation occurs with a corollary to the argument, that the human mind allegedly outperforms machines, because although there is no exhaustive formal definition of natural numbers, mathematicians can successfully work with natural numbers. Again, the corollary requires an extra-formal assumption: "**PA** is complete" or "the set of all natural numbers exists". I agree that extra-formal assumptions are necessary in order to validate the anti-mechanist argument and its corollary, and that those assumptions are problematic. However, I argue that formalization is possible and the problem is instead the circularity of reasoning that they cause. The human mind does not prove its own consistency, and outperforms the machine, simply by making the assumption "I am consistent". Starting from the analysis of circularity, I propose a way of thinking about the interplay between informal and formal in mathematics.

KEYWORDS: the Lucas-Penrose argument, the Church-Turing thesis, Carnapian explications, natural numbers, computation, conceptual engineering, conceptual fixed points, conceptual vicious circles, deviant encodings, structuralism.

[*] Warsaw University of Technology, Faculty of Administration and Social Sciences. International Center for Formal Ontology. E-mail: paula.quinon@pw.edu.pl. ORCID: 0000-0001-7574-6227.

1. Introduction

The Lucas-Penrose anti-mechanist argument against computability of the human mind in a nutshell states the following. According to Gödel's incompleteness theorems, a (sufficiently rich) consistent theory that can prove its own consistency does not exist. However, mathematical practice shows that Gödel-type results are commonly proven by human mathematicians. In consequence, says the argument, human mathematicians are not describable as formal proof systems, nor are they reducible to performing algorithms.

In (2020), Krajewski criticises the Lucas-Penrose argument by claiming that Gödel's incompleteness theorems standing alone (as it is in the Lucas-Penrose case) are not sufficient for formulating the claim that the human mind is non-computational. The anti-mechanist argument based on Gödel's incompleteness theorems needs to be enriched by an extra-formal assumption. For instance, an assumption that the theory constituting the human mind is consistent.

In order to provide an additional context to his investigations, Krajewski (2020), highlights the analogy between the claim that Gödel's incompleteness theorems imply the non-computational nature of the human mind, and the claim that "we [humans] cannot give a definition of the natural numbers as we understand them" (p. 49). The analogy goes as follows: in order to make a successful anti-mechanist argument based on Gödel's incompleteness theorems, one needs to assume—in addition to the formal counterpart—that the theory constituting the human mind is consistent. The fact that Gödel's argument can be iterated for increasingly rich theories is not sufficient for formulation of the anti-mechanist argument. The possibility to iterate increasingly rich theories, which all have a Gödel's sentence, and none of which proves its own consistency, is a formal process and as such can be executed by purely formal means. Thus, it does not say anything about computability or non-computability of the human mind. In order to be able to formulate the anti-mechanist argument, one needs to assume—for instance—that the human mind is consistent. Analogously, each definition of a natural number ends up in a vicious circle of definitions, or—as Krajewski says

> [O]ur axioms [both the first-order (**PA1**) and the second-order Peano Arithmetic (**PA2**)] define numbers only when taken together with some background knowledge or apparatus that makes possible our intuitive grasp of numbers [such as the intuition that the first-order Peano's Arithmetic is complete or the intuition that there exists the set of all natural numbers being referred to in the background of the second-order Peano's Arithmetic]. (2020, p. 49)

In both cases, an immediate, but incorrect according to Krajewski, conclusion could be that "no computer can be taught our concept of a number" and that in consequence "we [humans] are better than any machine" (2020, p. 49).

In this paper, I observe that this analogy can be pushed further to a circular reasoning. In both cases, making an extra-formal assumption leads to a vicious circle because one assumes consistency of one's mind while proving that the human mind outperforms machines, or one assumes that the concept of a set of natural numbers can be intuitively apprehended while defining natural numbers. Studies show that the method of conceptual analysis is particularly sensitive to falling into circular reasoning. The circularity related to the concept of natural number has been investigated in discussions about computational structuralism (Halbach & Horsten, 2005; Quinon & Zdanowski, 2007). Computational structuralism is a position, according to which the concept of natural number and the concept of computation are closely related. More precisely, according to this position, an adequate account of natural numbers treats them as objects that can be used for computations. After a brief overview of the anti-mechanist argument and its criticism in **Section 1**, in **Section 2** I will explain inter-relation and inter-definability between the concept of natural number and the concept of computation. In **Section 3**, I describe how the two concepts fall into a vicious circle of definition individually, and also while used in definition of one another.

Rescorla (2007) identifies problems with conceptual analysis related to the concept of computation, Quinon (2018) suggests that there is no fully satisfactory way out from vicious circles in definitions within conceptual analysis. Approaching the concept of computation and the concept of natural number from another methodological perspective, seems to be more fruitful. For instance, an interesting insight can be gained thanks to conceptual engineering. Both concepts have a form of what in the area of conceptual engineering is called "conceptual fixed point". A conceptual fixed point is an idea issued from the conceptual engineering of moral concepts, where it is claimed that some basic moral concepts should not be engineered, but should always be understood in the most objective way (Eklund, 2015). **Section 4** is devoted to the presentation of the method of conceptual engineering and the adequacy of conceptual fixed points for the concept of computation and the concept of natural number. As suggested by the phenomenon of conceptual fixed points, the only way out from these vicious circles consists in an arbitrary decision which is the intended meaning of the given concept.

In **Section 5**, I extend my methodological investigations into yet another method, and I discuss the advantages of thinking about formalisation of the concept of computation in terms of Carnapian explications. It has been argued, for instance in (Quinon, 2019), that a move from an intuitive concept of computation, used in everyday life, to a scientific or formal concept as stated by the Church-Turing thesis, follows the schema of a Carnapian explication. In **Section 6**, I extend the context of Carnapian explications of the temporary aspect. I realise that both, the concept of natural number and the concept of computation, have been evolving in such a way, that their core meanings were shifting. I propose a hypothesis that at least a part of the confusion regarding the specificity of the conceptual structure of the concept of computation contributes to the confusion regarding the nature of human reasoning and the human mind. In consequence, In consequence, I claim that—at least partially—the "feeling" that there are non-computational processes is due to the complexity of the conceptual structure of the concept of computation.

In the final **Section 7**, I wrap up with the ways in which my observations regarding the concept of computation and the concept of natural number, could be used for understanding the reasons for which the anti-mechanist argument fails. I suggest a different reason from the one proposed by Krajewski, for which the extra-formal assumption prevents the anti-mechanist argument from success. Firstly, I claim that thanks to the method of Carnapian explications, it is highly possible to go from intuitive pre-scientific concept to a formal concept. Secondly, I observe that the extra-formal assumption after an arbitrary formalisation, leads to the vicious circle in reasoning. Therein lies the problem.

2. The Lucas-Penrose Argument and Its Criticism

In this section, I present a brief overview of various versions of the anti-mechanist argument based on Gödel's incompleteness theorems, and the ways in which those arguments have been criticised. In particular, I explicate Krajewski's way of refuting the argument. In my overview, I prioritise the authors to who Krajewski refers to in his paper.

The first of Gödel's incompleteness theorem says that in every sufficiently rich[1] consistent first-order theory[2] there exist statements that are true[3], but that cannot be proven within this theory. The second of Gödel's incompleteness theorem says that every sufficiently rich consistent first-order theory cannot prove its own consistency.

According to the anti-mechanist argument based on Gödel's incompleteness theorems, since human mathematicians can fruitfully work with Gödel's incompleteness theorems, that means those mathematicians use the resources from the outside of the theory (e.g., they are able to refer to the intended model of arithmetic or recognize that the human mind is consistent). Thus, human mathematicians outperform machines, because—unlike machines—they are able to include in their reasoning such external resources.

The intuition that humans could prove theorems which machines could not has already been present in (Turing, 1950)[4] and in (Post, 1941).[5] One of the most famous voices exploring the anti-mechanist argument based on Gödel's incompleteness theorems against the computational theory of mind—next to Hofstadter

[1] By "sufficiently rich" one means that the formal system is able to express arithmetic of addition and multiplication.

[2] A formal system, or a theory, is a collection of axioms together with rules of inference. The importance of using first-order logic is because of the completeness of this logic.

[3] A statement is true, when it is satisfied in the intended model of the theory.

[4] As reported by Krajewski, Turing believed that even if a machine cannot prove as much as humans can, it is still worth constructing robots.

[5] As reported by Krajewski, Post believed that man cannot construct a machine which can do all the things he can.

(1979), Nagel and Newman (1958; 1961)—is Lucas (1961; also 1968; 1996), who presented a "mathematical proof" of man's superiority over a machine. Lucas extended the applicability of Gödel's incompleteness theorems from formal systems to human subjects. In his view, humans are subjects to the same formal limits as machines. However, as Lucas observes, human mathematicians can prove Gödel's incompleteness theorem, which means, human mathematicians use extra-formal resources that enable them to perform such proofs.

Lucas' argument relies on the fact that Gödel's theorem(s) is formulated in purely formal terms. As Lucas observes himself, this is what differentiates Gödel's results from the liar paradox. The liar paradox, which states that "This statement is untrue", is "viciously self-referential, and we do not know what the statement is, which is alleged to be untrue, until it has been made, and we cannot make it until we know what it is that is being alleged to be false" (Lucas, 1990, p. 2). Unlike the liar paradox, Gödel's theorem is formulated within a full-blooded system where it is clearly defined, which sentences are true and what does it mean to be provable. Lucas' claims that the fact that a (idealised) human mind, even if it cannot prove Gödel's theorem(s) for the given theory, can—thanks to its additional non-mechanical skills—recognize this theorem as true in its system. In consequence, a human mind outperforms a machine.

Penrose in (1989; 1994) extended Lucas' reasoning of a positive claim regarding the extra-formal resources available to humans that enable them to construct reasonings unavailable to machines. Penrose suggested that in the brain the physical basis of non-computable behavior exists, and he indicated quantum mechanics as a credible candidate. According to him quantum processes might explain not only reasoning of human mathematicians, but also consciousness.

A constructive criticism of the Lucas-Penrose style argument was formulated by Putnam (1960), Benacerraf (1967), Wang (1974), then later also by Boolos (1995) and Shapiro (1998). Penrose's version got criticised in particular by Feferman (1995), Putnam (1995) and Shapiro (2003). Krajewski claims that the ways of criticizing the Lucas-Penrose argument follow one of the two main lines (2020, pp. 5–6):

- The mind is a machine and it is consistent, but it cannot prove Gödel's sentence by itself.[6]
- The mind is a machine, but it is inconsistent, and Gödelian limitations do not apply to it.

[6] This line of argument has already come from Gödel, who distinguished *subjective arithmetic* that humans can do, and who believed that in *objective mathematics* full arithmetic is a consistent theory. He also believed that the concept of computation can be defined without referring to any domain of computation; these claims amount to Gödelian platonism (Gödel, *1951).

Krajewski (2020) refutes the Lucas-Putnam argument in yet another way: he observes that iterations of increasingly strong theories proving the corresponding Gödel's sentences can be processed in a purely mechanical or computational manner available to both, humans and machines. In consequence, Krajewski claims that anti-mechanist is not implied by Gödel's incompleteness theorems alone. In addition, claims Krajewski, one needs to assume that humans have a privileged access to assessing consistency of the human mind. Krajewski claims that the argument fails because of the necessity of making this extra-formal assumption. This is so, because there is no formal way to account for the formal counterpart of assumptions.

Before I come back, in the last section, to Krajewski's rejection of the anti-mechanist argument, and my proposal of how to shift the way of thinking about the reasons for this rejection, I will now focus on the part which is particularly interesting for me, that is the meta-theoretical corollary to the anti-mechanist argument stating that humans cannot fully describe the concept of natural number.

3. The Concept of Natural Number and the Concept of Computation

I initiate my investigation into the nature of the extra-formal elements of the reasoning that enable the conclusion that the human mind is not computable, by discussing the corollary relating human inability to define the concept of natural number. Additionally, I extend the corollary of the claim that humans—for similar reasons—cannot define the concept of computation. Finally, I present the view according to which the concept of natural number and the concept of computation are closely related.

The fact that every formal definition of the concept of natural number leads to a necessary assumption from the outside of the formal system has been studied in the context of the view in philosophy of mathematics, called structuralism. According to structuralism, mathematics is the "science of structures", and while defining mathematical objects, one should first target their structural properties. For instance, while defining natural numbers, one should define the structure of natural numbers through relations they hold to each other, and not focus on individual properties of those elements.

Traditionally, structuralism defined natural numbers using second-order Peano Arithmetic (**PA2**). **PA2** is categorical and the class of (isomorphic) models in which it is satisfied is identified with natural numbers. The usual way of criticising the use of second-order Peano Arithmetic to define natural numbers consists in saying that the underlying logic is "set theory in sheep's clothing" (Quine, 1970, p. 66). Second-order logic has the ability, for instance, to express the information that two sets have the same cardinality. The concept of set is itself most frequently (implicitly) defined with a first-order axiomatic theory, such as *ZF*, that in turn, is a subject of non-standard interpretations, the Löwenheim-Skolem theorem, etc., which makes its intended model "hidden" within a continuum of other non-intended models. Therefore, in order to define the concept of natural number with **PA2**, humans have two choices. They can get involved in a vicious circle of definitions, or an infinite regression of theorems, or they can use extra-formal resources and admit in an arbitrary manner that there is such a thing as an intended (or a standard) model of set theory where the intended model of arithmetic exists.

Another, less known, version of structuralism, so called *computational structuralism*, proposes distinguishing the standard model of arithmetic from the continuum of non-standard models with the resources of **PA1** only (Halbach & Horsten, 2005; Quinon & Zdanowski, 2007). In order to do that, defenders of computational structuralism suggest adding a meta-mathematical constraint regarding the computability of interpretation of functional symbols in the language, and then use Tennenbaum's theorem in order to single out the standard model of arithmetic.

Theorem 2.1 (Tennenbaum, 1959) *Let* $\mathcal{M} = \langle \mathbb{M}, +, \times, 0, 1, < \rangle$ *be an enumerable model of* **PA1***, and not isomorphic with the standard model* $\mathcal{N} = \langle \mathbb{N}, +, \times, 0, 1, < \rangle$*. Then* $\mathcal{M}$ *is not recursive.*

More explicitly why Tennenbaum's theorem is relevant for the structuralist way of thinking is visible in the transposition of the theorem:

Theorem 2.2 (Tennenbaum transposition) *Let* $\mathcal{M}$ *be an enumerable model of first-order Peano arithmetic. If the interpretation of addition and multiplication within* $\mathcal{M}$ *are computable then* $\mathcal{M}$ *is a standard model for arithmetic (a model with* ω*–type ordering).*

One of the philosophically interesting consequences of the application of Tennenbaum's theorem is that the set of models singled out with its help consists of those ω models, where ω is computable (Quinon & Zdanowski, 2007). Those models are called "intended" and form a proper subset of standard models.

The intended model of arithmetic,[7] is such a model where functions of addition and multiplication are interpreted as computable functions.[8] Tennenbaum's theorem establishes a connection between a meta-mathematical property of being computable by arithmetical functions, and the order of the elements of the set of natural numbers. Thus, in the most general lines, computational structuralism is a position, according to which the concept of natural number and the concept of computation are closely related.

The usual way of criticising computational structuralism is, again, by pointing out the vicious circle or infinite regression of definitions that threatens the proposed account of natural numbers. The criticism goes as follows: in order to

[7] Intended models of arithmetic are identified up to a computable isomorphism.

[8] The model of arithmetic is intended for both theories **PA1** and **PA2**.

define the concept of natural number, one needs to use the concept of computation, whereas every concept of computation is defined on the domain of (some representation of) natural numbers. Thus, the vicious circle or the necessity to assume that there is an intended interpretation of what to compute means, or that the intended model of arithmetic is distinguished from within other models.

Analogously, it is pretty straightforward that the concept of computation falls itself into a vicious circle, as in order to account for what "to compute" means, referring, for instance, "to be computed on a Turing Machine", necessitates to account for which entities are suitable for computing with (in the case of TM-computations, what can be the input for a Turing Machine). Since the question asked about the input precedes the definition of computing, which is just being given, one cannot use the concept of computing to define which sequences can be used for the input.

More precisely,

> [T]he Church-Turing Thesis states that Turing Machines formally explicate the intuitive concept of computability. The description of Turing Machines requires description of the notation used for the INPUT and for the OUTPUT. The notation used by Turing in the original account and also notations used in contemporary handbooks of computability all belong to the most known, common, widespread notations, such as standard Arabic notation for natural numbers, binary encoding of natural numbers or stroke notation. The choice is arbitrary and left unjustified. In fact, providing such a justification and providing a general definition of notations, which are acceptable for the process of computations, causes problems. This is so, because the comprehensive definition states that such a notation or encoding has to be computable. Yet, using the concept of computability in a definition of a notation, which will be further used in a definition of the concept of computability yields an obvious vicious circle. (Quinon, 2018, p. 338)

In this section, I explained similarities between the process of defining the concept of natural number, the process of accounting for the concept of computation, and the formulation of an anti-mechanist argument based on Gödel's incompleteness theorems. All these contexts are related because the way out of the definitional vicious circles proper to the definitional processes within formal theories, is through the necessity of assuming an additional non-formal, meta-theoretical knowledge. In the next section, I will expand on the phenomena of vicious circles and regression ad infinitum.

4. Nested Vicious Circles

Quinon (2018) proposes a taxonomy of what can be called "deviant encodings", that is those encodings—or in different words, sequences of symbolic representations of natural numbers—which are non-computable, but which are formally indistinguishable from computable encodings. For instance, in its simplest form the problem presents itself as follows:

> The problem in its purely syntactical version can be formulated as follows. In a definition of Turing computability, one of the aspects that needs to be clarified is the characterization of notation that can be used as an input for a machine to process. If a Turing Machine is supposed to explicate the intuitive concept of computability it is necessary to explain, which sequence of numerals can be used as an input without the use of the concept of computability. That means, we cannot simply say: "sequences that can be used as input are the computable ones" as we have not yet defined what it means "to be computable". (Quinon, 2018, p. 340)

Deviations refer to non-computable sequences that cannot be distinguished within the general formal context from sequences that are computable and can be used in computations. In this paper, I use the expression "deviant encoding" independently of the ontological framework within which natural numbers are understood. Quinon (2018) claims that the phenomenon of deviant encodings persists independently of which ontological status we assign to objects of computations (e.g., natural numbers, sequences of symbols, etc.). Quinon (2018) hypothesizes that the phenomenon of deviant encodings persists independently of the philosophical standpoint and provides an analysis of the following simplified standpoints: (i) purely mechanical/syntactical approach (nominalism, entwined mathematical concepts); (ii) notations have meanings (mild realism); (iii) semantics comes first (radical realism, platonic insight).

The study of conceptual "deviations" is conducted for a simplified framework where:

- on the syntactic level there are uninterpreted inscriptions, and where functions are string-theoretical generating string values from string arguments;
- on the semantic level there are interpretations that can range from the conceptual content ascribed to initially uninterpreted symbols, to Platonic abstract objects, and where functions are number-theoretical sending numbers to numbers;
- between the two levels there is defined a function of denotation.

Deviations occur on each level. Thus, there exist "deviant encodings" deviations that happen on the syntactic level; "deviant semantics" deviations that happen on the semantic level; "unacceptable denotation function" deviations of the denotation function.

The simplified framework is inspired by Shapiro (1982), who distinguishes string-theoretic functions from number-theoretic functions and searches for "acceptable", that is "non-deviant", ways of associating their domains. The framework is further used by other researchers. Rescorla (2007) uses it to study behaviour of denotation functions which associate numerals (symbolic representations of natural numbers) to natural numbers (abstract entities) in a non-computable manner. There is a continuum of such mappings.

The expression "deviant encodings" has been used differently by Copeland and Proudfoot (2010) for whom the deviations relate to encodings, or enumerations, of Turing Machines. The authors claim that a deviant encoding happens when the omniscient programmer "winks at us" to let us know when the number of a Turing Machine (from some standard encoding of Turing Machines), which is being currently processed by some sort of Halting Machine (a machine computing which Turing Machines stop on an input 0), refers to a machine that stops. In this way, the Halting Machine computes the halting function, which is an uncomputable function. The "wink" of the omniscient programmer gets encoded in the syntactic structure of the numerals: the numerals representing the machines that stop, have a special form—for instance—are even (their general syntactical form can be reduced to "2*n*" where "*n*" is any numeral). Copeland and Proudfoot mean by a deviant encoding such a standard enumeration of Turing Machines where the encoding is enriched by an extra-formal feature impersonated by the omniscient programmer (a Turing oracle). This is a specific case of a more general problem where deviant encodings refer to encodings representing natural numbers.

An occurrence of the phenomenon of deviant encodings involving all the levels, is the case of the Semantical Halting Problem (van Heuveln, 2000). Imagine, you have encoded Turing machines with some standard—computable, thus non-deviant—encoding, and that you believe that symbols have meanings or interpretations. It can happen that even if your syntax is generated in a recursive manner, your semantics is not following any recursive rules. The Halting Machine that processes encodings of Turing Machines is designed to process information on syntax in an algorithmic manner. If inputted with a given non-computable enumeration of Turing machines, the machine will process those non-computable encodings as if it were a standard notation. Again, there is no effective way of defining which semantics are acceptable and which are deviant.

I call "nested vicious circles" the hierarchies of vicious circles that keep reappearing at every stage of syntactical and semantic complexity of the presented picture.

To give an example of a philosophical position outside the strict theoretical context discussed in this paper, the phenomenon of deviant encodings appears as well in the case of concrete computations.

> In our ordinary discourse, we distinguish between physical systems that perform computations, such as computers and calculators, and physical systems that don't, such as rocks. Among computing devices, we distinguish between more and less powerful ones. These distinctions affect our behaviour: if a device is computationally more powerful than another, we pay more money for it. What grounds these distinctions? What is the principled difference, if there is one, between a rock and a calculator, or between a calculator and a computer? Answering these questions is more difficult that it may seem. (Piccinini, 2010)[9]

[9] See also Piccinini's (2015).

In (2020), Quinon notes that the phenomenon of nested vicious circles, relating to the concept of computability, does not disappear in the case of explicit inter-definiability between the concept of natural number and the concept of computation, as established by computational structuralism. As I have already described above, the criticism of computational structuralism consists in pointing at the choice between the definitional vicious circles or the necessity of making extra-formal arbitrary assumptions.

The way of extra-formal assumptions is investigated by Button and Smith (2012) who observed that when the concept "natural number" is explicated for, the concepts used in this explication, such as "to compute" or "finite" need to be accounted for on their turn, etc. In consequence, claim the authors, this problem cannot be tackled by offering more mathematics. An arbitrary decision regarding the meaning of some concept is necessary for the argument from Tennenbaum's theorem to work. However, as they claim in a slightly undermining way, this is a philosophical problem: "Suffice it to note that our discussion of Tennenbaum's Theorem illustrates a familiar moral: philosophical problems which are supposedly generated by mathematical results can rarely be tackled by offering more mathematics" (Button & Smith, 2012, p. 120).

Dean (2014) is similarly sceptical when it comes to the purposefulness of using Tennenbaum's theorem to formally single out the standard model of arithmetic. However, differently to Button and Smith, Dean develops a full-fledged philosophical position. It is a Putnam-style model-theoretic realism for the concept of computation (Putnam, 1980). Dean claims that there is no point in trying to find external arguments to distinguish between various standard and non-standard models neither of arithmetic, nor of recursive theory. We should rather use the richness of the model-theoretic universe for studying structural properties of the concept of computation. Dean claims that Tennenbaum's phenomenon shows that there exists a continuum of pairs: a model of arithmetic and computation in this model of arithmetic. In consequence, the Tennenbaum's result instead of contributing to singling out the standard model of arithmetic, it indicates that non-computable ω-models of arithmetic exist (the so called deviant or weird permutations) with a corresponding concept of computation defined within the model.

The vicious circle faced by computational structuralism, differs from the vicious circles that are the focus of Quinon (2018). There, I was only concerned by the concept of natural number being indirectly involved in the definition of what "to compute" means. Conceptual structuralism needs to handle a slightly more elaborate idea. Its objective is to explicate the concept of natural number, identified with the standard model of arithmetic. Its solution consists in using the idea that natural numbers, and in particular those which are defined by Peano's axioms, are the entities used for counting and computing. In consequence, natural numbers are defined in terms of computations. However, and this is where the vicious circle arises: one of the characteristic features of the concept of computation is that computation is always defined on some given domain.[10] This domain is always identifiable with the structure of natural numbers. I discuss the nested vicious circles in this context in (Quinon, 2020).

5. Conceptual Engineering and Conceptual Fixed Points

One of the promising ways out of the impasse consists in embracing that the circularity in the account of what "to be computable" and what "natural number" mean is due to limitations of conceptual analysis. Similarly to other scientific concepts, when analysis is conducted within the strict scope of a given formal theory, one often ends up with a necessity to use the concept which is being defined in the account of some concept used for its definition. Philosophers and logicians see in this feature of conceptual analysis both an advantage that enables us to understand more about the conceptual structure of the world (Dean, 2014), and a problem that blocks science from progress (Maddy, 2007). Rescorla (2007) identifies problems with conceptual analysis related to the concept of computation. In their paper (2012), Button and Smith claim that Tennenbaum's theorem is of no use to a philosopher who wants to distinguish the standard model from other possible models of arithmetic.

Quinon (2018) suggests that there is no fully satisfactory way out from vicious circles in definitions, resulting from conceptual analysis. Approaching the concept of computation and the concept of natural number from another methodological perspective, seems to be more fruitful. For instance, in recent years a particular type of conceptual work gained quite a bit of popularity, it is called *conceptual engineering*. What I try to convey in this section is that the new research on conceptual engineering actually provide additional insight into the possible ways of thinking about non-mathematical or non-formal knowledge.

According to Cappelen (2018), conceptual engineering is concerned with the assessment and improvement of concepts. As highlighted by Cappelen and Plunkett:

> since it's unclear and controversial what concepts are (and whether there are any), it's better to broaden the scope along the following lines:
>
> **Conceptual Engineering** = (i) The assessment of representational devices, (ii) reflections on and proposal for how to improve representational devices, and (iii) efforts to implement the proposed improvements. (2020, p. 3)

Researchers involved in developing the methodology of conceptual engineering realised that the method reaches its limits when concepts which are fundamental to the given theory are being scrutinised. They call it "conceptual fixed points". The most extensive reflection has been done in the area of ethics (Cap-

[10] A non-realised Gödel's objective consisted in finding an "absolute" concept of computation, *i.e.*, such a concept of computation that does not depend on any domain.

pelen et al., 2020), but Eklund (2015) extends it to formal contexts and concepts such as “truth”, “belief”, or “existence”. In addition to traditional arguments used in ethical contexts, such as “Kantian philosophy [with its regulative ideas], or from a naturalistic philosophy according to which what is innate severely constrains which concepts we can use”, Eklund considers basic formal concepts in the spirit of rigid designators.

In moral philosophy, “the moral fixed points” are those moral propositions that are moral truths which always need to be incorporated into a moral system. A normative system which fails to incorporate such propositions is not a moral system, but a normative system of some other kind. The leading example of such a moral fixed point is the proposition “It is wrong to engage in the recreational slaughter of a fellow person” (Cueno & Shafer-Landau, 2014).

Eklund (e.g., 2015, Chapter 5) extends this phenomenon to frameworks outside moral philosophy and, as he calls it, the “thinnest” normative words like “good”, “right”, “ought”. Eklund observes that in each conceptual framework, concepts exist that are difficult, if not impossible, to engineer. “Truth” is one of those concepts. People care about truth, writes Eklund, and they do not care about some conceptually engineered concept “truth*”. In consequence, truth is a concept that should keep a fixed position in a conceptual framework, and refer to the natural kin of assertions and beliefs. Similarly, “existence” is a conceptual fixed point. Eklund opposes the claim from the contemporary meta-ontological debate, where it is assumed “that there are alternative notions of existence that can be employed”. He claims that, similarly as in the case of “truth”, a conceptual framework that would result from adapting a conceptually engineered concept of “existence” would need to adjust its other key concepts in such a way that the resulting framework would be isomorphic to the initial one. Thus, “One cannot, so to speak, selectively engineer the quantifier”.

> Suppose we set out to conceptually engineer truth. Insofar as the job description of truth is that of being the property our beliefs and assertions aim at, the engineering project would be that of finding a property more adequate to that job description. But by what has been noted about Stich’s argument, it is hard even properly to conceive of a practice of belief or assertion that is guided by a different property. (Eklund, 2015, p. 378)

There is one last thing that I consider worth mentioning while talking about conceptual fixed points and mathematical concepts, in particular the concept of computation, that is a possible proximity between conceptual fixed points and fixed points that are traditionally analysed in mathematics in the context of diagonalisation. At first sight, they do not have much in common[11] as conceptual fixed points relate mostly to the cross-model intended interpretation of a concept, whereas diagonalisation is about self-reference and vicious circles. Conceptual fixed points are concepts interpreted in, what we call in the philosophy of math-

[11] I might be wrong, but I will not try to sort it out in this paper.

ematics, their intended models. In different words, a fixed point consists of the pair the engineered concept corresponding to the intended meaning of the concept, or—to borrow Eklund's expression—the interpretation that "people care about", and a possible world of interpretation, which actually corresponds to the intended model of this concept. Both, the concept of natural number and the concept of computation are in this sense conceptual fixed points. A more careful look should be applied to those two phenomena, but in this paper I will just leave it without further comment.[12]

6. The Church-Turing Thesis as a Carnapian Explication

Another methodological framework that offers a solution for conceptual structure escaping conceptual analysis is the method of Carnapian explication. Quinon (2019) explores the idea that the structure of the concept of computation, accounted for with the Church-Turing thesis, is best understood through the method of explication. This section is devoted to the presentation of the method of explication for the concept of computation, and also for the concept of natural number.

Treating the concept of computation, as accounted for in the Church-Turing thesis, as a Carnapian explication has multiple advantages, namely, it overcomes problems of conceptual analysis; it explains how one intuitive concept of what "to be computable" means can be translated into a multitude of extensionally equivalent formal concepts of "to be computable" in a specific formal concept means; it finally provides a ground for thinking of mathematical or formal concepts as "open-textures" evolving through time (Makovec & Shapiro, 2019); it also relates to the initial intuitive prescientific concept with the formal concept, because an explication relies on an existing meaning, and offers a specification which offers the best possible fit in a given context.

An explication in the Carnapian sense consists in introducing new formal concepts to the scientific language coined on the basis of everyday concepts. In different words, it is a procedure of transformation from an inexact prescientific concept into a scientific one. Moreover, an explication consists in providing a scientific concept within a given context, within an existing theory. It is done in two steps:

- The clarification of the explicatum
- The specification of the explicatum

The rationale for clarification is that a given term may have many different meanings in ordinary language. Unless one of these meanings is clearly picked

[12] If you want to get a more formal description of this phenomenon, you can think of hybrid modal logics which provide a framework for thinking of epistemic access to other possible worlds from the perspective of the selected distinguished world.

out from the start and the context of its use is clearly indicated, it is unlikely that the method of explication will yield a useful result. Clarification serves this purpose. As Carnap explains, “[a]lthough the explicandum cannot be given in exact terms, it should be made as clear as possible by informal explanations and examples” (Carnap, 1950, p. 3). Quinon (2019) highlights the importance of the clarification stage, the stage which has traditionally been underestimated.

A clarification of the explicandum enables the next step of the explication process, a specification of the explicatum and formulation of the exact concept in the targeted context.

Since several clarifications most often can be foreseen, and several scientific contexts are available, one pre-scientific concept can be explicated in various manners. In order to decide which explication is the most successful, Carnap proposes four criteria that can be applied for assessing the value of an explication, and also for comparison between available options.

- SIMILARITY TO THE EXPLICANDUM: most of the cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.
- EXACTNESS: the rules of use of explicatum have to be given explicitly and precisely, for example, by providing a concept with the formal definition.
- FRUITFULNESS: shall be “useful for the formulation of many universal statements”.
- SIMPLICITY: an explication should be as simple as the previous three allow it.

I think that it is worth investigating whether abandoning the path of analysis and taking the path of explications could offer an additional insight into the conceptual structure of formal concepts, and also informal concepts lying in the foundations of their formalization. The idea is that every formal concept is—at least in subjective arithmetic (to borrow Gödelian terminology)—grounded upon, or issued from, an everyday intuitive, pre-scientific concept. The next section is devoted to a preliminary investigation into the possibility of extending the idea that the method of explication, consisting in building up the formal concept out of the intuitive concept, is anyhow relevant to the anti-mechanist argument against the computability of mind using Gödel’s incompleteness theorems.

Both intended interpretations determined in the consequences of accepting conceptual fixed points solution and the choice of the formal aspect, and the formal context at the stage of the concept clarification in the process of Carnapian explication, share a similar threat. In the case of a fixed point solution and in the case of clarification an agent needs to take an arbitrary decision regarding the intended interpretation.

7. Theory of Mind and Computations

In this section, I propose an additional complication to the method of Carnapian explication, which is a temporary, or a phylogenic, aspect of conceptual development.

The method of Carnapian explication enables introducing new formal concepts to the language by transforming an intuitive pre-scientific concept into a new scientific concept within some formal context. Usually, at the stage of clarification one chooses the meaning that will guide the formalisation of the intuitive pre-scientific concept and also the targeted formal context. What I propose in this section, is an additional dimension to the clarification stage: a relativisation to the phylogeny of the formal concept. At the stage of clarification, in addition to deciding which aspect of the intuitive concept one wants to formalise, one needs to realise that each concept develops. The phylogenic development of the concept of natural number and the concept of computation is studied in Shapiro on open-texture (2013).

The relation between the concept of computation and the concept of natural number underwent a very dynamic development. In consequence, the set of potential clarifications of intuitive concepts of computation and of natural numbers have grown. What is interesting from my perspective, is that computability is today an expected feature of natural numbers. Natural numbers are those mathematical entities that are all day long used for enumerating and computing, for programming, and in various sorts of logistic projects as an underlying discrete structure. Both concepts have become increasingly important in the everyday life of our society. This is called digitalisation.

Various areas of digitalisation are additionally reinforced by the fact that computationalism—even if its formal details are still discussed by philosophers, mathematicians and logicians—is today the mainstream theory of mind. This process is described by Turkle (1984; 2011; 2015) who studies how concepts from computer sciences and robotics have got into common language and how they have changed ordinary people's approach to inter-personal relations or ethical questions.

According to Turkle the intensity in which digitalisation of everyday life develops is strongly connected to the fact that computational language was first used to reformulate our perception of our own mind and consciousness.[13]

[13] Turkle's earlier work related to a similar development of conceptual trends in explanation of phenomena of everyday life that had a place in France in the 1960s and 1970s as a consequence of the spread of psychoanalytical ideas, see her book *Psychoanalytic Politics: Jacques Lacan and Freud's French Revolution* from 1978). In *The Second Self: Computers and the Human Spirit* (1984), Turkle describes these changes that have got into general culture through digitalisation and robotics in the same way as "psychoanalytic culture" penetrated structures of the general social and political life in France: "Psychoanalytic language spread into the rhetoric of political parties, into training programs for schoolteachers, into advice-to-the-lovelorn columns. I became fascinated with how people were picking up and trying on this new language for thinking about the self. I had gone to

When Turkle speaks about her experience with the digitalised society, she compares two experiences:

> My experience at MIT impressed me with the fact that something analogous to the development of a psychoanalytic culture was going on in the worlds around computation. At MIT I heard computational metaphors used to think about politics, education, social process, and, most central to the analogy with psychoanalysis, about the self. (Turkle, 1984, p. 305)

She sees within it a first step in the cultural assimilation of a new way of thinking:

> The essential question in such work is how ideas developed in the world of high science are appropriated by the culture at large. In the case of psychoanalysis, how do Freudian ideas move out to touch the lives of people who have never visited a psychoanalyst, people who are not even particularly interested in psychoanalysis as a theory? In the study of the nascent computer culture, the essential question was the same: how were computational ideas moving out into everyday life? (Turkle, 1984, p. 305)

She searches how "the idea of mind as a program enters into people's sense of who is the actor when they act". A model of the mind that is adapted by society influences how people think about their frustrations and disappointments, their relationships with their families and with their work (Makovec & Shapiro, 2019, p. 305). On the other hand, says Turkle, computers became a new constructed object—"a cultural object that different people and groups of people can apprehend with very different descriptions and invest with very different attributes. Ideas about computers become easily charged with personal and cultural meanings" (Turkle, 1984, p. 308).

In her other books, Turkle studies human attachment to objects. In the volume of essays *Evocative Objects: Things We Think With* (2007) she speaks about the attachment that people, many of her friends, have developed with physical objects. In her book, *Alone Together* Turkle (2011) extends her observations to different types of automated artificial agents, such as virtual agents mediated by electronic support, or robots. In a series of social experiments, where she asked her subjects to interact with an automated artificial agent, she observed that the stronger attachment develops in the most vulnerable members of our society, such as neglected children with unfulfilled emotional needs, or with old people suffering from a lack of human interactions. Our natural inclination to form emotional attachment with humans, and with objects in the absence of humans, might soon lead to even more human-AI interactions. Those interactions are obviously

France to study the psychoanalytic community and how it had rein- vented Freud for the French taste, but I was there at a time when it was possible to watch a small psychoanalytic community grow into a larger psychoanalytic culture" (Turkle, 1984, pp. 304–305).

structured in a very particular, very automated, way, which even more strongly influences the digitalisation of the language we use.

Krajewski makes a similar observation in the last section of the paper.

> Our attitude toward the arguments of Lucas, Penrose, and others is shaped mostly by our general vision of machines and minds. And this vision adjusts with changes of civilization. For the youth of today, if I may judge from listening to my students, our computerized world makes it easier to accept the idea that anything is mechanizable—including the mind. (2020, p. 49)

I propose a hypothesis that at least part of the confusion regarding the specificity of the conceptual structure of the concept of computation contributes to the confusion regarding the nature of human reasoning and the human mind. In consequence, I claim that—at least partially—the "feeling" that there are non-computational processes is due to the complexity of the conceptual structure of the concept of computation.

8. The Lucas-Penrose Argument and Extra-Formal Concepts

Let me now come back to the anti-mechanist argument against computability of mind based on Gödel's incompleteness theorems.

In the first part of this section, I reconstruct Krajewski's claim according to which, in order to make the anti-mechanist argument work, one needs to add an extra-formal assumption stating the consistency of the underlying theory, that is, the theory corresponding to the human mind. The core of Krajewski's criticism is as follow: it is not possible to formalize the extra-formal assumption and therefore, the whole of Lucas' argument is fallacious. I disagree with Krajewski's claim that formalization of the extra-formal assumptions is not possible. There are contemporary philosophical methods that might enable formulation of such a formalization. As example, in the previous sections, I have presented the methodological and conceptual framework was based on Carnapian explications. Instead, I focus on another problem, which the issues from an internal characteristic of formal contexts, namely on the part of the argument, which leads to a circular reasoning. In order to show that the human mind (T_{HM}) outperforms a machine (T_M), one needs to assume that the human mind is consistent and knows it (and in this way outperforms a machine that can never "know" if it is consistent or not). Observe, that I do not reject Krajewski's conclusion, but I point at a fallacy in a proof. Again, I have already discussed how the method of conceptual engineering enables structured thinking of extra-formal assumptions and the resulting circular reasoning.

In the second part of this section, I will continue my investigation of possible extra-formal assumptions relative to the anti-mechanist argument based on Gödel's incompleteness theorems.

The Lucas' anti-mechanist argument based on Gödel's incompleteness theorems consists of two parts. Firstly, Gödel's results establish that each sufficiently rich consistent theory admits a Gödel sentence and also that none such theory can prove its own consistency.

Let T be a consistent theory containing arithmetic, let φ_T be the Gödel's sentence for the theory T.

$$Con(T) \rightarrow T \nvdash \varphi_T$$
$$Con(T) \rightarrow T \nvdash Con(T)$$

Moreover, it is broadly known that an inconsistent theory proves any sentence, but Gödel's incompleteness theorems do not apply to an inconsistent theory.

Secondly, human mathematicians can work with subsequent increasingly stronger theories,

$$T_1 = T \cup Con(T)$$
$$T_2 = T_1 \cup Con(T_1)$$
$$\vdots$$
$$T_{n+1} = T_n \cup Con(T_n)$$

which—for some defenders of the anti-mechanist argument—signifies that human mathematicians outperform machines. Krajewski objects to this view claiming that the construction of the hierarchy can be fully mechanised. In consequence, he claims that the ability to construct and work with the hierarchy of increasingly stronger theories alone is not sufficient for formulating the anti-mechanist argument. As stated by Krajewski, additional assumptions are missing.

> In addition to Gödel's results, at least two assumptions that are not self-evident are used in the above reasoning. First, every exact proof of our consistency can be formalized, second, it is possible to express "our consistency". […] If this is accepted, one could question the second point. It is not clear at all how one can express "our consistency". Basically there are two options to express this: either (i) by the common sense statement "I am consistent" or (ii) by a formal counterpart to this statement. Let us consider them in turn.
>
> In case (i) we refer to a common sense statement, which have no connection to formal considerations. Hao Wang (1974, pp. 317–320) reflected on just this statement and believed that it is not provable. […] If that were possible, it would mean that we are not machines, or that we are not even equivalent to machines in the realm of proof-producing reasoning. We certainly may believe that, but it is no more than a general feeling.
>
> In case (ii) we consider the formal counterpart to a loose statement expressing consistency […]. The usual meaning of the statement refers to the will to avoid contradictions, to the reliability of our vision of the world, to the claim that the methods used by mathematicians are unfailing. The sentence *Cons* or any other similar arithmetical formula is rather far from those ideas. Thus, while something is strictly proved, it is unclear to what extent the conclusion conveys our consistency. (2020, pp. 47–48)

Krajewski's reasoning can be reconstructed as follows. Applying the formal predicate "being consistent" can only apply to a formal theory. Applying the formal predicate "being consistent" to anything else than a formal theory is a categorical mistake. In consequence, if "consistency" is to be a predicate applying to on the human mind, the mind must have certain formal properties and needs to be identified with a theory. The following options exist:

- If human mind is a theory and it is consistent, then as to all other theories, a Gödel's sentence applies to it and the human mind encounters the same constraints as any theory (a machine).
- If the human mind is a theory and it is inconsistent, then Gödelian argument limitations do not apply at all.

If the human mind is a theory, a human disposing of a mind cannot know—from the formal point of view—if it is consistent or not. In consequence, in order to prove that the human mind outperforms a machine, a second extra-formal additional assumption needs to be made. It has to be assumed that the human mind is indeed consistent. This assumption can be done in one of the two ways. "Case (i)", "I am consistent" cannot be formalised. "Case (ii)", there exists a formal counterpart of "I am consistent".

My analysis of "case (i)" is in line with the analysis of Krajewski. If "I am consistent" is an informal statement, it is useless for any formal proof. And here we speak of being able to prove more than a machine. Whereas Lucas' argument is supposed to be a formal proof of the superiority of the human mind over a machine.

My analysis of "case (ii)" differs from Krajewski's analysis. His argument returns to the idea that each formalisation of the informal "I am consistent" remains—maybe more informed or more precise—but is still an informal account. As such it is useless for any formal proof. I think that the conclusion from (ii) is different. An agent can find a formal counterpart of the statement "I am consistent", or rather "the theory constituting my mind is consistent". The framework of the Carnapian explications enables us to understand how it can be done.

I also assume that an agent can recognise their own consistency. This insight is available to a human being, while it is—on the grounds of the second of Gödel's incompleteness theorem—unavailable to a machine. This extra-formal assumption is necessary for formulating an anti-mechanist argument against the computability of the mind. It is also exactly at this point where a vicious circle occurs. We are in the act of proving that the human mind outperforms a machine, and so one cannot in this proof assume that human mind is consistent.

Another possible extra-formal assumption that can be made in order to enable the anti-mechanist argument based on Gödel's incompleteness theorem, is the ability to refer to the intended model of arithmetic.[14] Instead of assuming that the human mind is consistent (i.e., assuming that the theory underlying all human reasoning is a consistent theory, which does not prove both a φ and a $\neg\varphi$, for every φ), in order to use Gödel's incompleteness theorems to support the anti-mechanist argument, one can assume that the human mind is able to refer to the intended model of arithmetic. The assumption that the human mind can refer to the intended model of arithmetic disables the possibility that the Gödel sentences get to have non-standard Gödel numerals.

In the way it is usually interpreted—in particular in the context of philosophical argumentation supporting the anti-mechanist argument that the human mind is non-computable—Gödel's incompleteness theorems provide us with the information from the perspective of a formal system. The semantical aspect is taken for granted. When the model-theoretical reasoning is applied, Gödel's incompleteness theorems indicate that there exist non-standard models in which the (non-standard) Gödel number of the proof for Gödel's incompleteness theorems has its (semantical) reference. It also means, that there exist models where the Gödel (non-standard) number of the proof for the negation of Gödel's first theorem, has an interpretation as a (non-standard) natural number.

What is famously referred to by Gödel's platonism is his belief that there is a model of arithmetic in which all arithmetical truths are satisfied. This is obviously not the intended model of arithmetic that humans have privileged cognitive access to, but the model of arithmetic in objective mathematics (Gödel, *1951).

9. Conclusions

Additionally to the critical analysis of Krajewski's rejection of the anti-mechanist based on Gödel's incompleteness theorems to which I suggest some possible improvements, my paper is sympathetic to the idea that certain key concepts in formal contexts naturally fall into circular or infinite reasonings. In this way, I try to shift attention from the theory of the human mind and consciousness, to the study of the conceptual structure of the language.

In my paper, I explored similarities between various formal contexts in which key concepts fall into a vicious circle of reasoning. I looked at the formalisation of the concept of natural number, of the concept of computation, and at the concept of consistency in the context of Gödel's incompleteness theorems. I suggested that the way to switch from an informal pre-scientific concept to a full-blooded formal scientific concept formulated in an adequate formal context is best modeled by Carnapian explications. I have also suggested that the phenom-

[14] The intended model is intended for both **PA1** and **PA2** and for this reason I do not make a distinction between the intended model of **PA1** and the intended model of **PA2**. I can think of a philosophical position that makes such a distinction, but for my purpose that would unnecessarily complicate my presentation.

enon of conceptual fixed points offers a methodological framework to think of intended interpretations necessary to jump out of circularity.

REFERENCES

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

Boolos, G. (1995). Introductory Note to *1951. In: S. Feferman et al. (Eds.), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 290–304). Oxford University Press.

Button, T., Smith, P. (2012): The Philosophical Significance of Tennenbaum's Theorem. *Philosophia Mathematica*, *20*(1), 114–121.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press.

Cappelen H., Plunkett D. & Burgess A. (Eds.). (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Carnap, R. (1950). *Logical Foundations of Probability*. Routledge and Kegan Paul.

Copeland, J., Proudfoot, D. (2010). Deviant Encodings and Turing's Analysis of Computability. *Studies in History and Philosophy of Science*, *41*, 247–252.

Cuneo T., Shafer-Landau, R. (2014). The Moral Fixed Points: New Directions for Moral Nonnaturalism. *Philosophical Studies*, *171*, 399–443.

Dean, W. (2014), Models and Computability. *Philosophia Mathematica*, *22*(2), 143–166.

Eklund, M. (2015). Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In C. Daly (Ed.), *The Palgrave Handbook of Philosophical Methods* (pp. 363–385). London: Palgrave Macmillan.

Feferman, S. (1995). Penrose's Gödelian Argument. *Psyche: An Interdisciplinary Journal of Research on Consciousness*, *2*, 21–32.

Gödel, K. (193?), Undecidable Diophantine Propositions. In S. Feferman et al. (Eds), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 164–175). Oxford University Press.

Gödel, K. (*1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications [Gödel's 1951 Gibbs lecture]. In S. Feferman et al. (Eds.), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 304–323), Oxford University Press.

Halbach, V., Horsten, L. (2005). Computational Structuralism. *Philosophia Mathematica*, *13*(2), 174–186.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach, and Eternal Golden Braid*. New York: Basic Books.

Krajewski, S. (2007). On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in any Attempt to 'Out-Gödel' the Mechanist. *Fundamenta Informaticae*, *81*, 173–181.

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

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

Lucas, J. R. (1968). Satan Stultified: A Rejoinder to Paul Benacerraf. *The Monist*, *52*, 145–158.

Lucas, J. R. (1990). A Paper to Read to the Turing Conference at Brighton on April 6th, 1990. Retrieved from: http://users.ox.ac.uk/~jrlucas/Godel/brighton.html

Lucas, J. R. (1996). Minds, Machines and Gödel: A Retrospect. In P. Millican, A. Clark (Eds.), *Machines and Though* (pp. 103–124). Oxford University Press.

Maddy P. (2007). *Second Philosophy. A Naturalistic Method*. Oxford University Press.

Makovec, D., Shapiro S. (Eds.). (2019). *Friedrich Waismann. The Open Texture of Analytic Philosophy*. New York: Springer.

Nagel, E., Newman J. R. (1958). *Gödel's Proof*. New York University Press.

Nagel, E., Newman J. R. (1961). Answer to Putnam. *Philosophy of Science*, *28*, 209–211.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Piccinini, G. (2010). Computation in Physical Systems. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press.

Plunkett D., Cappelen, H. (2020). A Guided Tour of Conceptual Engineering and Conceptual Ethics. In: H. Cappelen, D. Plunkett, A. Burgess (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp. 1–26). Oxford University Press.

Post, E. (1941). Absolutely Unsolvable Problems and Relatively Undecidable Propositions—Account of an Anticipation. In M. Davis (Ed.), *The Undecidable* (pp. 338–433). Hewlett, N. Y.: Raven Press.

Putnam, H. (1960). Minds and Machines. In S. Hook (Ed.), *Dimensions of Mind: A Symposium* (pp. 138–164). New York: New York University Press.

Putnam, H. (1980). Models and Reality. *Journal of Symbolic Logic*, *45*(3), 464–482.

Putnam, H. (1995). Review of The Shadows of the Mind. *Bulletin of the American Mathematical Society*, *32*(2), 370–373.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press.

Quinon, P. & Zdanowski, K. (2007). Intended Model of Arithmetic. Argument from Tennenbaum's Theorem. In S. B. Cooper et al. (Eds.), *Computation and Logic in the Real World* (pp. 313–317). Berlin: Springer-Verlag.

Quinon, P. (2014). From Computability Over Strings of Characters to Natural Numbers. In A. Olszewski, B. Brożek, P. Urbańczyk (Eds.), *Church's Thesis, Logic, Mind & Nature* (pp. 310– 330). Warsaw: Copernicus Center Press.

Quinon, P. (2018). Taxonomy of Deviant Encodings. In: F. Manea, R. Miller, D. Nowotka (Eds.), *Sailing Routes in the World of Computation* (pp. 338–348). Berlin: Springer-Verlag.

Quinon, P. (2019). Can Church's Thesis be Viewed as a Carnapian Explication? *Synthese*, Online First.

Quinon, P. (2020). Implicit and Explicit Examples of the Phenomenon of Deviant Encodings. *Studies in Logic, Grammar and Rhetoric*, *63*(76), 53–68.

Rescrola, M. (2007), Church's Thesis and the Conceptual Analysis of Computability. *Notre Dame Journal of Formal Logic*, *48*(2), 253–280.

Shapiro, S. (1982). Acceptable Notation. *Notre Dame Journal of Formal Logic*, *23*(1), 14–20.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *Journal of Philosophical Logic*, *4*, 273–302.

Shapiro, S. (2003). Mechanism, Truth, and Penrose's New Argument. *Journal of Philosophical Logic*, *32*, 19–42.

Shapiro, S. (2013). Computability, Proof and Open-texture. In A. Olszewski, J. Wolenski, R. Janusz (Eds.), *Church's Thesis After 70 Years* (pp. 420–455). Berlin: Walter de Gruyter.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460.

Turkle, S. (1978). *Psychoanalytic Politics: Jacques Lacan and Freud's French Revolution*. New York: Basic Books.

Turkle, S. (1984). *The Second Self: The Second Self: Computers and the Human Spirit*. MIT Press.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Turkle, S. (2015). *Reclaiming Conversation: The Power of Talk in a Digital Age*. London: Penguin Press.

van Heuveln, B. (2000). *Emergence and Consciousness: Explorations Into the Philosophy of Mind via the Philosophy of Computation* [Unpublished Ph.D. thesis]. State University of New York, Binghampton.

Wang, H. (1974). *From Mathematics to Philosophy*. Routledge and Kegan Paul.