1. Errata

- p. xiii, line 20: instead of “Metalanguage” read “English”.
- p. 198, line −10: instead of “v_M(x ∈ y)” read “v_M(y(x)).”
- p. 198, line −9: instead of “the formula x ∈ y” read “the sentence y(x).”
- p. 201: lines −9 and −13: instead of “↔” read “←.”
- p. 202, lines 1: instead of “if and only if” read “if.”
- p. 203, line 26: instead of “two variables” read “3 variables”.

2. Commentaries

2.1. Dogmatism. Our presentation of logic and mathematics is largely dogmatic in that it ignores alternative approaches and does not justify the (apparently) arbitrary (meta)definitions and axioms that are being introduced. Dogmatism definitely helped keeping our introduction minimal. But the question of a priori justifying our (meta)definitions and axioms remains to be addressed. The quest for such a priori justifications is, in some sense, analogous to Hegel’s requirement (Hegel 1975) that the “list of categories”, for instance, be a priori deduced rather than “dogmatically given”; Hegel criticizes Kant for precisely such a dogmatic approach.

In our case, for instance, one can ask: what is the justification for having exactly 5 connectives ∧, ∨, ¬, →, ↔ and what is the justification for their particular “truth tables”? One answer could be that our starting point were natural languages which we started analyzing before we synthesized our more formal languages. And, in analyzing natural languages, we just happened to empirically find 5 connectives that are being used in such a way that, if actual truth were an allowed predicate, then the “truth tables” would coincide with the corresponding actual truth tables. But this justification is not acceptable. First it is an a posteriori, rather than an a priori, justification. Secondly this justification is at odds with our insistence that the truth predicate be banned from our discourse; our “truth tables” are, rather, part of the inference machine. Here is, however, an a priori justification, in line with our ban on truth. There are 16 possible “truth tables” for a binary connective and 4 possible “truth tables” for unary connectives. So there are 16 + 4 = 20 possible binary+unary connectives. So there are 15 “new” connectives in addition to the “old” 5 that we have been using. (We can call them △, □, ..., etc. and we can even translate these 15 connectives into English as, say, gook, gonk, etc.) Now it turns out that any of the 15 new connectives is expressible as a Boolean combination in terms of the old 5 connectives (and indeed in terms of, say, ¬ and ∧ alone); this is an easy exercise. So, after all, the old 5 connectives are enough to define all 20 possible (unary+binary) connectives. This somewhat dispels the mystery of the 5 connectives and their special “truth tables”; for they “generate” all possible connectives, i.e. all possible “truth tables”. It would be interesting to find a similar
a priori justification for the background axioms, the specific axioms of set theory (ZFC), and the various definitions within mathematics. The ZFC axioms are an especially puzzling case: their shape is far from intuitive and hence far from easily justifiable a priori.

2.2. Plurality of logics and mathematics. One of the corollaries of our dogmatic approach is that only one pre-mathematical logic was considered; and within it only one mathematics (namely ZFC) was put forward; and within this particular mathematics only one mathematical logic was set up. But lacking an a priori justification for the choices made in developing the theory one should point out towards various alternatives. Indeed consideration of a pre-mathematical logic involves setting up a set of syntactic rules; let $S_1, S_2, \ldots$ be the various possible sets of syntactic rules. Belonging to each set $S_n$ of syntactic rules there are various languages $L_{n1}, L_{n2}, \ldots$. For each $n$ one can select a language, say $L_{n1}$, that we want to use for mathematics. In this language $L_{n1}$ one can formulate various sets $A_{n11}, A_{n12}, \ldots$ of axioms (which are variants of “axioms for mathematics”). For each $n$ and $m$ we can then give various metadefinitions $D_{n1m1}, D_{n1m2}, \ldots$ of the concept of theory. All of these lead to corresponding theories $T_{n1m1}, T_{n1m2}, \ldots$ (which are variants of “mathematics”). Within each $T_{n1mk}$ one can set up various systems of definitions $\Delta_{n1mk1}, \Delta_{n1mk2}, \ldots$ each of which allows us to create a mirror of $T_{n1mk}$ inside itself; for each $\Delta_{n1mk}$ we get a different “mathematical logic.” Then the syntactic rules we put forward in Part 1 of our book identify with one of our sets of syntactic rules, say with $S_1$. The language $L_{set}$ introduced in Part 2 of our book identifies then with $L_{11}$. The ZFC axioms identify with one of the possible sets of axioms, say with $A_{111}$. Our specific metadefinition of the concept of theory corresponds to some $D_{1111}$, say. Set theory $T_{set}$ identifies then with $T_{1111}$. And the definitions used to set up mathematical logic in Part 3 of the book identify, say, with $\Delta_{11111}$. So one can see that, in setting up pre-mathematical logic, mathematics, and mathematical logic, one actually has a whole spectrum of choices at each point. This perspective was fully understood and thoroughly explored by the classics of contemporary logic, in particular by Gödel who made sustained efforts to enrich ZFC (i.e. enrich $A_{111}$) in such a way that the continuum hypothesis $CH$ can be “decided” within the enriched set theory. Gödel also looked carefully at how his incompleteness theorems change if one “weakens” $L_{11}$ and $A_{111}$. He noted, cf. (Wang 1996), that his arguments for the incompleteness theorem of arithmetic are not finitistic, as Hilbert had required for “metamathematics”. In fact, it was pointed out by Gödel that his arguments of the incompleteness of arithmetic require the full power of the language/axioms of set theory and not merely the language/axioms of the integers; in other words, for the proof of these theorems, one cannot substantially weaken $L_{11}$ and $A_{111}$. This was viewed by Gödel as an argument in favor of the “metaphysical reality” of sets.

2.3. Metaphysics. Traditional metaphysics (from Aristotle to Medieval to Leibniz) is concerned with problems such as: “being qua being” (existence in itself), “ontology” (existence of the world, of substances, what exists, how many substances there are), “first causes” (theology, rationality of the world, “logos”), “universals” (abstract terms, the unchanging), etc. Modern metaphysics (especially post Kant) has put more stress on problems of epistemology (sensibility, understanding, knowledge, the dichotomies of a priori/a posteriori and analytic/synthetic, the nature of
space/time/causality), problems of consciousness (the mind/body problem, free will), problems of psychology, etc. Our book attempts a presentation of logic, and in particular of mathematics, that is as free as possible from both traditional and modern metaphysics. This is mostly done by ignoring the metaphysical problems that are inherent in logic (as in the theory of meaning, reference, truth) and more generally in linguistics. These problems are real; but, as shown by our book, their solution does not seem to be required for a presentation of mathematics. Remarkably, Gödel believed (along Leibniz’s suggestions and contrary to Kant’s critique) that traditional metaphysics can be made into an exact science; here is a quote from Gödel reproduced in (Wang 1996): “Philosophy as an exact theory should do for metaphysics as much as Newton did for physics.” Again, according to Wang, Gödel believed that complicated abstract concepts (such as sets, which he viewed as concepts) can be perceived as clearly as simpler abstract concepts (such as syntactic combinatorics, which he also viewed as a system of concepts); he also seems to have confessed that he could not eventually find a satisfactory system of primitive notions for metaphysics.

2.4. Mock metaphysics. One can turn the tables and view (traditional) metaphysics as part of logic by introducing a language $L_{\text{metaphys}}$ whose constants are existence, world, God, this or that object, etc., whose variables are substance, predicate, etc., and whose predicates are exists, predicates, is one, is infinite, etc. This is doable but leads to a “mock metaphysics” which is a “copy” of metaphysics inside logic. One can iterate this move and create a copy of the “mock metaphysics” (a “mock mock metaphysics”) inside mathematics (i.e. inside set theory) in the same way in which logic itself has a copy inside mathematics which is mathematical logic. Now recall Alain Badiou’s maxim that the metaphysics of existence qua existence is nothing but [Cantor] set theory. (By set theory Badiou seems to understand Cantor set theory or any metaphysically realist version of set theory.) This could then be transposed inside mathematics by saying instead that, “the mock mock metaphysics [that we discussed above] is a chapter of the set theory $T_{\text{set}}$” (where now set theory $T_{\text{set}}$ is in the sense of our book, in particular $T_{\text{set}}$ is just a text). The moral of the above discussion is that all these mock-ifications of metaphysics are not metaphysics itself and say nothing about metaphysics itself (in the same way in which the theorems of mathematical logic say nothing about logic itself, simply because, being sentences in object language, they say nothing at all). The only way “mock mock metaphysics” can be relevant to metaphysics is to have its sentences translated into a natural language such as English. (In the same way mathematical logic can only be relevant to mathematics after translation into English, say.) Such translations are suspect and easily lead to contradictions; this is why we made it a rule in our book to ignore translations from object languages. On the other hand if metaphysics is to exist at all it has to be a prerequisite of (or to incorporate) logic. Our book, of course, is trying to avoid making metaphysics a prerequisite of logic, though, and this seems to work if the project is modest enough. Introducing mathematics seems to be one of these modest projects.

2.5. Physics. In our book the term is used in the modern (rather than Aristotelian) sense; it is physics as a scientific body of knowledge, as it emerged after Galileo and Newton. It is written in a language containing abstract terms (force,
mass, energy) and it comes with translations into both mathematics (set theory) and the language of empirical data (statements about measurements).

2.6. Realism, conceptualism, nominalism. According to (Quine 1980) the three doctrines in the 20th century philosophy of mathematics (logicism, intuitionism, and formalism) correspond to the three “medieval points of view regarding universals designated by historians as realism, conceptualism, and nominalism.” Realism holds that “universals or abstract entities have being independently of the mind.” (Putnam calls this type of realism metaphysical realism or externalism.) Conceptualism holds that “there are universals but they are mind-made.” And nominalism objects to “admitting abstract entities at all.” For the nominalist all that exists in connection with an abstract entity such as set, or number, or force, or mass, or God, is the corresponding word (as a physical mark on paper). Nominalism in science and mathematics (under its variant referred to as logical positivism) was the position put forward by the Vienna Circle (most notably by Carnap); Gödel himself, although part of the Vienna Circle at one point, rejected nominalism and made a case, later in his life, for a realist position. One can arguably further subdivide nominalism into two types which we could call $T$ and $NT$; type $T$ recognizes objective truth as a meaningful concept (mostly through a correspondence theory) whereas $NT$ (which is a rather rare and radical variety) denies the meaningfulness of objective truth. Our book is written from the viewpoint of an $NT$ type nominalism. The classical books of Quine (Quine 1980) and Putnam (Putnam 1971, Putnam 1981) reject nominalism based on their commitment to a concept of “objective truth.” Quine says (Quine 1980, p. 121) that he is adopting a “liberal ontology” (admitting classes, hence sets) while Putnam says (Putnam 1971, p. 23) that “reference to ‘classes’ [...] is indispensable to the science of logic.” In particular Putnam makes a strong case against what we referred to as type $T$ nominalism. However he does not consider type $NT$ nominalism as a possibility and his arguments against type $T$ do not seem to apply to type $NT$. There is one caveat in the assertion above that our approach is from the viewpoint of an $NT$ nominalism; namely, although sentences in a language can be viewed as concrete objects (signs on a piece of paper), the theories $T, T', \ldots$ considered in the book may be accused of being non-nominalist entities (because, for instance, they are potentially infinite entities whose very consideration seems to imply an abstract conception). This accusation has some merit and the way the book implicitly responds to it is as follows. The more abstract our discussion about theories becomes (i.e. the more one starts using words such as consistency, completeness, etc.) the less is one allowed to metaprove metasentences involved in such discussion. In the limit, when discussion about theories becomes utterly abstract, the discussion transforms itself into “un-metaprovable babbling.” This is the very reason why the book defers all matters related to consistency and completeness to mathematical logic and essentially bans them from pre-mathematical logic. What saves the day here is the ban on truth predicates true/false. The “babbling” referred to above is neither true nor false.

2.7. Variation on an argument of Mach. Here is an argument against realism in mathematics. Assume one adopts a correspondence theory of truth in mathematics (based on a realist position) and one considers a sentence such as, “For any positive integer $n$ there is a prime bigger than $n$.” In order to ascertain the truth
of such a sentence one needs to “check it” against “reality”. Assuming the “reality” of the integers the only way to check the above is to perform an experiment. Since we assume the actual infinity of the integers the only possible type of experiment is a “thought experiment”: one fixes, mentally, a positive integer \( n \) and one puts forward a procedure that constructs a prime bigger than \( n \). But an objection can be raised similar to Mach’s objection to thought experiments in physics. Recall that Mach’s objection (originally addressing Newton’s “bucket experiment”) is that it is meaningless to talk about the outcome of a thought experiment if the conditions for the practical implementation of that experiment are inherently “impossible” to achieve. One should not accept thought experiments, for instance, that allow the size of the experimenter’s apparatus to be the size of the Earth itself because there is no way to know what would happen if this were the case. In the same way we do not know if the procedure to find a prime bigger than a given \( n \) would function as expected in case \( n \) is a number bigger, say, than the “number of atoms in the universe.” It is conceivable that the “rules of logic” do not apply as expected to such an \( n \). Similarly a thought experiment that starts with an “impossible” premise such as, “Assume the whole universe, except the Earth, were to disappear...,” should not be accepted; and in a similar vein mathematical proofs by contradiction (that usually start with an “impossible” premise as well) should not be accepted from a realist standpoint.

2.8. Variation on a maxim of Badiou. In the preface to his Logics of Worlds (Bloomsbury, London, New York, 2009) Alain Badiou says that, “today, natural belief is condensed in a single statement: There are only bodies and languages.” He then proposes to replace the 2 by 3 by introducing a “universal exception” as follows: “There are only bodies and languages, except there are truths.” What our approach amounts to is to the replace the 2 by 1 by introducing another “universal exception” as follows: There are only bodies and languages, except all bodies are made of language. The latter point of view is sufficient for pure mathematics but it is not appropriate, of course, for more ambitious philosophical endeavors. Indeed ontology is a matter of commitment and commitment is a pragmatic attitude. In the book we say that the only things we accept as existent are the signs on paper that can be assembled into languages (according to rules that are also expressed within yet another language). By this we simply mean that, when dealing with mathematics, we choose to ignore everything else except language. This is a commitment based on a practical decision to develop a certain specific project. Such a position leaves room for more elaborate ontologies oriented towards more ambitious projects. One motivation for a more inclusive ontology would be that, with a minimal ontology that only recognizes the reality of language, there is no room for non-linguistic reference. Discussion of non-linguistic reference may become relevant, however, when we ask the following question: What does mathematical logic say about mathematics itself? Cf. 2.24 below.

2.9. Variation on the field metaphor of Quine. Here is an example of a more inclusive ontology that can be used to accommodate the interaction between mathematics and natural sciences. It is the ontology implicit in Chapters 11 and 43 and to present it we shall elaborate on a metaphor of Quine’s (cf. Quine 2008, p. 43) using some Kantian jargon. One can image a sphere the things outside of which are called the noumenal world. The things inside the sphere are called
symbols. Both the noumenal world and the symbolic world are viewed as empirical/physical: the symbols are for instance written or spoken; we just choose to view them as irreducible/simple entities rather than complexes of images, sounds, etc. Symbols are articulated into systems called languages. The articulation itself is encoded into other languages and hence into other systems of symbols. There are various dynamical reports between the various languages (translations, reference, etc.) which are encoded into yet another language, etc. The noumenal world acts like a *boundary condition* (in a sense similar to that encountered in the theory of partial differential equations). The symbolic world acts like a *field*. The field adjusts itself according to the boundary conditions; in particular logic itself is part of the field and adjusts itself to the noumenal world. The law according to which this adjustment takes place is one that seeks equilibrium (i.e. simplicity/intelligibility) very much like Maupertuis’ law of minimum action. But this law does not correspond to a “noumenal law”; it is rather a law obeyed by the “symbolic world”. The surface of the sphere is the *phenomenal* world. The strings of symbols close to the surface are *concepts*. The strings of symbols closer to the center of the sphere are the *ideas*. There is no direct relation (in particular semantic relation) between specific things outside the sphere such as the “real” planets and specific things inside the sphere such as the word *force*: the way the outside influences the inside is purely global (as in the case of boundary conditions influencing a field). The above metaphor is still “nominalist” but its ontology is more generous than the one adopted in this book. An interesting question related to this metaphor is: what are the “field equations”? Can they be written in one of the languages inside the sphere? This corresponds, in the metaphor above, to one of the main projects of critical philosophy.

### 2.10. Balance between object language and metalanguage.

The fine tuning of object language and metalanguage is based on the following “balance” principle best expressed in a table as follows:

<table>
<thead>
<tr>
<th></th>
<th>object language</th>
<th>metalanguage</th>
</tr>
</thead>
<tbody>
<tr>
<td>syntactic structure</td>
<td>strong</td>
<td>weak</td>
</tr>
<tr>
<td>semantic structure</td>
<td>weak</td>
<td>strong</td>
</tr>
<tr>
<td>ability to infer</td>
<td>strong</td>
<td>weak</td>
</tr>
<tr>
<td>ability to refer</td>
<td>weak</td>
<td>strong</td>
</tr>
<tr>
<td>truth</td>
<td>banned</td>
<td>banned</td>
</tr>
</tbody>
</table>

### 2.11. Intension and extension.

Classically the *intension* of a predicate in a language is its “dictionary” definition. So what we call *definition* of a new predicate in our book (e.g. *an astrochicken is a thing which is both a chicken and a spaceship*) corresponds to the intension of the predicate *is an astrochicken*. On the other hand the *extension* of a predicate $P$ is, classically, the collection of all “objects in the world” that can be predicated by $P$, i.e. of which $P$ is “true.” Since “truth” is banned from our book there is nothing in the book that corresponds to extension of predicates; see the discussion below on *truth and extension* for more details on this. Finally there is a third way to fix predicates in a language, namely by *ostention*, i.e. by pointing, e.g. by pointing to an astrochicken and saying, “*this is an astrochicken*”. As the very example we gave shows, ostention leads to a very
limited array of definitions; in particular it cannot be used to define astrochickens simply because there seems to be none that one can point at.

2.12. Interplay between truth and extension. Our ban on the predicate “true” (and its synonyms such as “is the case”) prevents one from considering extensions of predicates. To see this assume one is given an object language \( L \) and a Cantor set \( M \) (so \( M \) is a collection of actual things called elements) equipped with some relations

\[ R \subseteq M \times M \times \ldots \]

(so that sentences such as \((x, y, \ldots) \in R\) are either objectively true or objectively false). Let us call \( M \) a realist model of \( L \) if a rule is given that attaches to the constants of \( L \) elements of \( M \) and to the predicates of \( L \) relations on \( M \). (Realist models have little to do with the metamodels or the models in the book.) Philosopher also call \( M \) a possible world and they call the rule an interpretation. Given a predicate \( P \) in the object language \( L \) one classically calls the extension of \( P \) in the realist model \( M \) the collection of all elements of \( M \) for which \( P \) is objectively true; so if truth is banned from our discourse so will be extension of \( P \) in \( M \). The axiom in ZFC that postulates, for any \( y \), the existence of a set \( \{ x \in y; P(x) \} \) has, strictly speaking, nothing to do with the concept of extension in \( L_{\text{set}} \); and this is not only because of the presence of \( y \) but mainly because that axiom in ZFC, being a sentence in an object language, has no semantics! Here we recall that sets in \( T_{\text{set}} \) are not collections of things but rather constants in \( L_{\text{set}} \), hence marks on a piece of paper. We stress the fact that the discussion above only applies to object languages; in particular it does not apply to predicates in metalanguage \( \hat{L} \). Indeed if we consider a system of axioms and we consider the predicate is a theorem in \( \hat{L} \) then the extension of this predicate is the theory \( T \) generated by the given system of axioms! Here we recall that \( T \), in particular \( T_{\text{set}} \), is a collection; this is in contrast with sets which are constants in \( L_{\text{set}} \) and hence are not collections.

2.13. Comment on Tarski’s scheme. In a realist model theory (that is based on the maximal ontology of Cantor sets or on any other similar realist ontology), one classically says that a sentence \( P \) in an object language is true in \( M \) if the sentence becomes true as a sentence “about the elements of \( M \).” This classical definition of truth in a (realist) model (essentially due to Tarski) does not solve the problem of defining truth; it just defines truth in \( L \) in terms of truth about Cantor (or other realist) sets. But truth about Cantor sets is left undefined! (As a matter of fact the whole Tarski scheme can be reproduced inside mathematical logic, as opposed to pre-mathematical logic; this is being discussed in Chapter 47. The result is a “formal” version of Tarski’s scheme. But being formal, this version has no semantic content, hence no direct report with the non-formal version of Tarski’s scheme.) In any case our ban on truth predicates puts Tarski’s theory of truth outside the scope of our book.

2.14. Truth and reference in Wittgenstein’s Tractatus. Wittgenstein’s Tractatus Logico-Philosophicus seems to be based on the concept of “is the case”, as synonymous of “is true,” and is, therefore, entirely perpendicular to the viewpoint of our book. The very first sentence of the Tractatus reads, “The world is all that is the case.” The Tractatus then offers a vision of reality as a tightly articulated logical structure in which all sentences are either the case or not the case. Nothing is more remote from the picture put forward in our book where objective truth
is treated as metaphysics and banned from discourse. By the way, in relation to reference, the Tractatus says, “There must be something identical in a picture and what it depicts, to enable the one to be a picture of the other at all.” (Wittgenstein, 2.161). This amounts to an attempt at a logical proof of the fact that reference is based on similarity of form.

2.15. Similarity and reference. Similarity theory of reference is, of course, as old as philosophy, and was thoroughly criticized by empiricists like Locke and, at a deeper level by Kant and all the resulting German idealist tradition, down to the internalist project of Putnam, say. For an example showing that similarity does not imply reference see (Putnam 1981); Putnam’s example is that of an ant tracing a line in the sand that is a recognizable caricature of Winston Churchill: here we have similarity without reference. And of course reference does not imply similarity as one can see in the example of the “three feet long table” whose mental image is not three feet long and indeed has no length at all.

2.16. Predicativity, impredicativity. When adding definitions of new constants, functions, or predicates to a language one should ask, of course, that these definitions be introduced in a sequence and at each step in the sequence the symbol that is being introduced has not appeared before (i.e. it is indeed “new”); this guarantees the predicativity of the definitions (i.e., essentially their non-circularity), at least from a syntactical viewpoint. This device does not get rid of the semantic impredicativity (which was one of the major themes in the controversies around the foundation of mathematics at the beginning of the 20th century; cf. Russell, Poincaré, and many others.) However, since we chose to completely ignore the meaning (semantics) of object languages, semantic impredicativity will not be an issue for us. To be sure, later in set theory, semantic impredicativity is everywhere implicit and might be viewed as implicitly threatening the whole edifice of mathematics.

2.17. Cross-reference. Introducing new definitions in a language (be it object language or metalanguage) is a form of cross-reference; cross-reference acts as synonymy internal to the language and, as such, it is different from what, in our book, we call reference (the latter being always reference to things external to the language such as another language, as in linguistic reference to the “word Napoleon”, or even noumenal objects such as the “man Napoleon”. ) There are other forms of cross-reference; the very use of variables (or, from a grammatical viewpoint, pronouns) plays the role of cross-reference. On the other hand synonymy between fragments of different languages is covered by our concept of translation, and hence, by semantics.

2.18. Indeterminacy of meaning. In our presentation the meaning of a sentence is defined to be the totality of all available translations of the sentence. The catch word here is “available”; it implies that the translations are viewed as given, or rather that they are the output of a “black box” which we are ignoring. But of course, for the philosophy of language, this black box itself if the interesting thing. The problem that needs to be answered is: how to define meaning other than “by fiat”? In other words: if meaning is defined in terms of translations then how are translations (e.g. dictionaries, or encyclopedias) possible if meaning is not yet present? For an illuminating discussion of these issues see (Quine 1980), Chapter
III, The problem of meaning in linguistics, and (Quine 1964). One of the main ideas in loc.cit. is that meaning is underdetermined (i.e. there is no way to fix it through the concept of logical truth); this is, in some sense, compatible with our “black box” metaphor. Cf. especially Quine’s careful discussion of radical translation (i.e. translation from/to an unknown language) in (Quine 1964), in particular his description of what would take to create a Jungle-to-English dictionary (where Jungle is the language of a remote tribe none of whose speakers speak English). The creation of such a dictionary would require a detailed interaction with the members of the tribe such as sequences of questions that would identify their words for yes and no as well as their words for particular objects such as rabbits. This dictionary work would actually not guarantee that the Jungle word that we translate as rabbit in English actually refers to rabbit and does not refer to a part of the rabbit or an instance of rabbit-hood, etc. Quine’s discussion is mostly relevant to natural (as opposed to artificial/formal) languages. For the (artificial) language of set theory (i.e. of mathematics) the issue of meaning can be, in principle, completely avoided. Nevertheless, the metalanguage needed to operate with the language of set theory is treated in our course as a natural language and its own meaning (that was not discussed in the book) requires further clarification.

2.19. Translation and possible worlds. Note that among translations of an object language into other object languages one finds what we called earlier the interpretations in possible worlds (or equivalently the realist models). The latter are, in today’s philosophy of language, the point of departure of semantics (e.g. of concepts such as extension); so our working definition of meaning is consistent with (although more minimalist, hence more vague than) the accepted definition of meaning.

2.20. Reference versus meaning. This issue goes back to Frege’s breakthrough in the philosophy of language. Frege’s famous example is: the “morning star” and the “evening star” have different meanings but, as was discovered at some point, they turn out to have the same reference (the same physical star); if meaning and reference were the same the previous sentence would be tautological, which it is not, because it accounts for a discovery, so meaning and reference are different. In our book references to things such as stars are ignored because we ignore non-linguistic reference. However linguistic reference is being considered in the form of metalanguage referring to language. Also meaning is being considered in the form of translations between object languages (e.g., as in the case of metamodels). The main difference between reference and meaning comes, in our book, from the fact that translations (in their simplest word for word form) attach constants to constants, predicates to predicates, etc., (hence they preserve logical categories) while linguistic reference attaches sentences, and even sequences of sentences, in object languages to constants in metalanguage, connectives and quantifiers in object language to functional symbols in metalanguage, etc. (hence it shifts logical categories). In the case of not-word-for-word translations logical categories are also shifted but only in the small and never in the large; for instance sentences in one language are always sent into sentences. Also we postulate that translations preserve reference; i.e. if L and L' have a reference and if P is a sentence in L whose translation in L' is P' then P and P' have the same reference.
2.21. Philogeny of reference, meaning, categories. This is mostly an anthropological (rather than philosophical) problem. One modern answer (Cas-sirer 1951) involves the insights of the history of mythical thought, from its first stage of “momentary” deities (perceived as spontaneous apparitions accompanying concrete objects such as this or that tree) to more permanent deities (functioning as names for abstract concepts such as the general concept of storm, crop, etc.). Another modern answer (put forward by many including Bronowski, rejected by many, including Putnam) involves natural selection. The meaning of the concept of “cause”, for instance, could have been fixed as follows. Say that hominid A throws a rock at hominid B and, as a result hominid B dies. Hominid C witnesses the scene and his mental apparatus produces the following description of the events: “The cause of death of hominid B was A’s intention to kill B with a rock”. Let us also assume that another hominid, D, witnesses the same scene and his mental apparatus produces the following description of the events: “The cause of death of hominid B was the property of B’s head to attract rocks present in A’s hand”. Now hominid C will survive by being careful to avoid hominids such as A; then C will have offspring who will inherit his particular mental wiring. Hominid D, on the other hand, will probably die soon at the hands of the likes of A; he will have no offspring and his type of mental wiring will not be inherited by anybody. After generations most hominids will have a mental wiring similar to C’s; the category of “causality” with its more or less fixed meaning will have emerged in this way. This would explain not only the meaning and reference of the abstract term “causality” but also the origin of the a priori form of understanding (in Kant’s sense) of “causality”. Whether or not such an account has merit is beyond our scope here.

2.22. Ontogeny of reference, meaning, categories. The problem is how reference, meaning, categories are fixed in a given individual. There are rival theories, for which we refer, for instance, to (Quine 1964) and (Chomsky 2006) respectively.

2.23. Impossibility of translation into metalanguage. Recall from 2.20 that we postulate that translation preserves reference i.e. if L and L′ have a reference and if P is a sentence in L whose translation in L′ is P′ then P and P′ have the same reference. We claim now the following:

Assume
1) L has a reference;
2) L is not self-referential (i.e. no sentence of L refers to sentences of L);
3) Ḳ is a metalanguage for L (i.e. it refers to L);
Then no translation from L into Ḳ is possible.

Indeed here is a metaproof of this. Assume there is a translation from L into Ḳ and take a sentence P of L whose translation in Ḳ is  ḲP. Since Ḳ refers to L we have that ḲP refers to some sentences Q, R, ... in L. But P and ḲP have the same reference because translations preserve reference. So P refers to Q, R, ... But then L is self-referential, a contradiction. The above should qualify as a metaproof (at least if we view all letters as constants).

2.24. What does mathematical logic say about mathematics? Since Gödel’s theorems in Chapter 48 are sentences in L^s, they have, according to the conventions in the present course, no meaning and no truth value. In particular
these theorems say nothing about mathematics itself. One could ask, however, what happens if we declare at this point that $L_{set}$ actually has a reference (which we ignored so far but we are ready now to take into consideration) and that $L_{set}$ is not self-referential (which we have to be careful to insist on); is it the case, under these new assumptions, that Gödel’s theorems do actually say something about mathematics itself? We will give an argument below that, even under these new assumptions, Gödel’s theorems still cannot be seen (in our paradigm) as saying anything about mathematics itself. Indeed let us take the following Theorem of Gödel (a strengthening in $T_{set}$ of the corresponding one in the text) as an example; it is the following sentence in $T_{set}$:

$$P_{con} \rightarrow (P_{con} \notin \Theta_{set}).$$

It can be reformulated as the following sentence in $L_{set}$:

$$(0.1) (\Theta_{set} \neq \Lambda_{set}^*) \rightarrow ((\Theta_{set}, \neq, \Lambda_{set}^*) \notin \Theta_{set}).$$

A statement about mathematics itself (i.e. about $T_{set}, L_{set}$, etc.) which would be of interest and that seems to correspond to 0.1 would be the following metasentence in the metalanguage $\hat{L}_{set}$ (that talks about $L_{set}$):

$$(0.2) \text{If } T_{set} \neq L_{set} \text{ then } "T_{set} \neq L_{set}" \text{ is not in } T_{set}.\$$

Now all we seem to have to do is try to consider 0.2 as a translation of 0.1. However this is fallacious: indeed, under our assumptions, there is no translation from $L_{set}$ into $\hat{L}_{set}$! Cf. 2.23. We reached the conclusion that,

\textit{Even if we allow the language $L_{set}$ of mathematics to have a (non-linguistic) reference, Gödel’s theorems (and indeed mathematical logic as a whole) cannot say anything about mathematics itself.}

Of course this conclusion was a corollary of our formalist (nominalist) position. On the other hand Gödel himself was a Platonist (metaphysical realist) and his own conclusions with regards to what his theorems are saying about mathematics were rather different from the one expressed in this course. For more on Gödel’s Platonist views see (Cohen 2008) and (Wang 1996).

\textbf{2.25. The ontological argument.} This could be added as an example in Chapter 14. The structure of the classical ontological argument for the existence of God is as follows. Let us assume that qualities (same as properties) are either positive or negative (and none is both). Let us think of \textit{existence} as having 2 kinds: \textit{existence in mind} (which shall be referred to as \textit{belonging to mind}) and \textit{existence in reality} (which shall be referred to as \textit{belonging to reality}). It is not important that we do not know what mind and reality are; we just see them as English words here. The 2 kinds are not necessarily related: belonging to mind does not imply (and is not implied by) belonging to reality. (In particular we do \textit{not} view mind necessarily as part of reality which we should not: unicorns belong to mind but not to reality.) The constants and variables refer to things (myself, my cat, God,...) or qualities (red, omnipresent, deceiving, eternal, murderous, mortal,...); we identify the latter with their extensions which are, again, things (the Red, the Omnipresent, the Deceiving, the Eternal, the Murderous, the Mortal,...) In particular we consider the following constants: reality, mind, God, the Positive Qualities. We also consider the binary predicate \textit{belongs to}. We say a thing has a certain quality (e.g. my cat is...
eternal) if that thing belongs to the extension of that quality (e.g. my cat belongs to the Eternal). Assume the following axioms:

A1) There exists a thing belonging to mind that has all the positive qualities and no negative quality. (Call it God.)

A2) Belonging to reality is a positive quality.

A3) Two things belonging to mind that have exactly the same qualities are identical.

(Axiom A3 is Leibniz’s famous principle of identity of indiscernibles. It implies that God is unique.) Then one can prove the following:

**Theorem 0.1.** God belongs to reality.

The above sentences are written in the English language $L’$. Let us formalize the above in a language $L$ and prove a formal version of Theorem 0.1 in $L$ whose translation is Theorem 0.1. Assume $L$ contains among its constants the constants $r, m, p$ and a relational binary predicate $E$. We consider a translation of $L$ into $L’$ such that

- $r$ is translated as “reality”;
- $m$ is translated as “mind”;
- $p$ is translated as “the positive qualities”;
- $xEy$ is translated as “$x$ belongs to $y$”.

The specific axioms are:

A1) $\exists x((xEm) \land (\forall z((zEp) \leftrightarrow xEz)))$.
A2) $rEp$.
A3) $\forall x\forall y(((xEm) \land (yEm)) \rightarrow ((\forall z(xEz \leftrightarrow yEz)) \rightarrow (x = y)))$.

Note that later, in set theory, we will have a predicate $\in$ which, like $E$, will be translated as “belongs to”; but the axioms of $E$ and $\in$ are very different. Let us make the following:

**Definition 0.2.** $g = c((xEm) \land (\forall z((zEp) \leftrightarrow (xEz))))$.

So $g$ is defined to be equal to a certain existential witness. Note that $gEm$ is a theorem by A1. We will translate $g$ as “God”. We have the following theorem expressing the uniqueness of God:

**Theorem 0.3.** $\forall x(((xEm) \land (\forall z((zEp) \leftrightarrow (xEz))) \rightarrow (x = g)))$.

The translation of the above in English is: “If something in my mind has all the positive qualities and no negative quality then that thing is God.”

**Proof.** A trivial exercise using axiom A3 only.

On the other hand, and more importantly, we have the following Theorem whose translation in $L’$ is “God belongs to reality”:

**Theorem 0.4.** $gEr$.

**Proof.** By axiom A1 and the quantifier axioms we have $gEm$ and

$\forall z((zEp) \leftrightarrow (gEz))$.

Hence, again by the quantifier axioms, we have, in particular,

$$(rEp) \leftrightarrow (gEr).$$

By axiom A2, $rEp$. Hence, by modus ponens, we have $gEr$. 

□
The argument above is, of course, correct. What is questionable is the choice of the axioms and the reference of \( L \). Also recall that, in our book, the question of truth was not addressed; so it does not make sense to ask whether the English sentence “\textit{God has existence in reality}” is true or false. For criticism of the relevance of this argument (or similar ones) see, for instance, (Kant 1991) and (Wang 1996). However, the mere fact that some of the most distinguished logicians of all times (in particular Leibniz and Gödel) took this argument seriously shows that the argument has merit and, in particular, cannot be dismissed on trivial grounds.

2.26. Dialectic. First order logic, which is the basis for mathematics, fixes its concepts in terms of definitions and axioms. And these are viewed as “outside time” (“eternal”). One can wonder where dialectic (Socratic, Hegelian, etc.) fits into this picture. Dialectic, identified with the identity game that is being played between contraries, does play a role in first order logic but only at the stage where one sets up the definitions and axioms of a theory (cf. the discussion in Wang 1996); once axioms and definitions are fixed the resulting theories are viewed as “frozen for eternity” and immune to dialectic.

3. Alternative versions of Chapters 47 and 48

Below is a alternative version of Chapters 47 and 48. In this alternative version the material is slightly rearranged, more details are given, and some of the abuses of notation in the original treatment are being avoided. One of the effects of the rearrangement is that the theory \( T'_{set} \) was eliminated from the main body of the presentation and is only mentioned now in a final remark. What is being achieved is a presentation which is safer \( (T'_{set} \) has a higher chance than \( T_{set} \) of being inconsistent); the price to be payed is to not allow “proofs that there is a formal proof of \( P' \)” to count as “proofs of \( P \)”. The introduction of the symbol \( \dagger \) in the alternative version below is meant to eliminate one of these abuses of notation from the original presentation; what is being achieved is more conceptual clarity but the price to be payed is, of course, having to deal with slightly heavier notation.
Chapter 47: Models

(Alternative version)

We briefly indicate here how one can create a “mirror” of pre-mathematical logic within mathematics (indeed within algebra). What results is a subject called mathematical logic (also referred to as formal logic). This mirroring process can be thought of as a “second formalization” (or a formalization of the formalization). A special place in mathematical logic is played by models/model theory which we shall explain below in some detail. In particular we will discuss models of ZFC itself (and hence of mathematics itself); this concept of model is entirely different from that of metamodels in the chapter on metamodels. In the discussion of models of ZFC we will introduce what one can call formalized set theory. We will discuss models of other (simpler) theories as well such as the theory of Peano arithmetic.

The mirror of pre-mathematical logic in mathematics is not entirely accurate: there is no one to one correspondence between pre-mathematical logic and mathematical logic. But as a general principle if a concept (say crocodile) is present in pre-mathematical logic its mirror in mathematical logic will acquire the adjective formal in front (e.g., will be called a formal crocodile). The dichotomy non-formal/formal indicates the difference between set theory in pre-mathematical logic and what we shall call formalized set theory in mathematical logic. There is another dichotomy, the syntactic/semantic dichotomy; semantic means, roughly, coming from translation of sentences whereas syntactic means coming from the shape of the sentences.

All definitions and theorems below are in set theory $T_{set}$. Recall that $\mathbb{N}, 0, 1, 2, 3, ...$ are sets, i.e. constants in $L_{set}$, that were defined in the chapter on the integers.

Definition 0.5. For $i \in \mathbb{N}$ define sets

$$c, f_i, r_i, x_i^\dagger, \land^\dagger, \lor^\dagger, \neg^\dagger, \rightarrow^\dagger, \leftrightarrow^\dagger, \forall^\dagger, \exists^\dagger, =^\dagger, (^{\dagger})^\dagger$$

as follows:

- $c = 0$,
- $f_i = (1, i)$,
- $r_i = (2, i)$,
- $x_i^\dagger = (3, i)$,
- $\land^\dagger = 1$,
- $\lor^\dagger = 2$,
- $\neg^\dagger = 3$,
- $\rightarrow^\dagger = 4$,
- $\leftrightarrow^\dagger = 5$,
$\forall \dagger = 6,$
$\exists \dagger = 7,$
$= \dagger = 8,$
$(\dagger, =) = 9,$
$I = \{c\} \cup \{f_1, f_2, f_3, \ldots\} \cup \{r_1, r_2, r_3, \ldots\},$
$V = \{x_1, x_2, x_3, \ldots\},$
$W = \{\forall \dagger, \exists \dagger, \neg \dagger, \rightarrow \dagger, \leftrightarrow \dagger, \forall \dagger, \exists \dagger, = \dagger, (\dagger, \dagger)\}.$

$V$ is called the set of variables; $W$ is called the set of logical symbols. We sometimes write $x^\dagger, y^\dagger, z^\dagger, \ldots$ instead of $x_1, x_2, x_3, \ldots$. By a $T$-partitioned set we mean in what follows a set $S$ together with a map $S \rightarrow T$. We let $S_t \subset S$ the preimage of $t \in T$. Let $S$ be a $T$-partitioned set; the elements of $S_t$ are called constant symbols; the elements of $S_{\mathfrak{r}_n}$ are called $n$-ary function symbols; the elements of $S_{\mathfrak{r}_n}$ are called $n$-ary relation symbols. For any such $S$ we consider the set $\Lambda_S = V \cup W \cup S$

(remarked to as the formal language attached to $S$). Then one considers the set of words $\Lambda_S$ with letters in $\Lambda_S$. One defines (in an obvious way, imitating the metadefinitions in the chapters on “pre-mathematical” logic) what it means for an element $\varphi \in \Lambda_S$ to be an $S$-formula or an $S$-formula without free variables (the latter are referred to as sentences). One denotes by $\Lambda^S \subset \Lambda_S$ the set of all $S$-formulas and by $\Lambda^S \subset \Lambda^T$ the set of all $S$-sentences.

Remark 0.6. Note that symbols “$\land, \lor, \ldots$” are, of course, connectives in $L_{\text{set}}$ while “$\forall \dagger, \exists \dagger, \ldots$” are sets, i.e. constants in $L_{\text{set}}$. Also “$=$” is equality in $L_{\text{set}}$ while “$= \dagger$” is a constant in $L_{\text{set}}$. Also “$(,)$” are separators in $L_{\text{set}}$ while “$(\dagger, \dagger)$” are constants in $L_{\text{set}}$. Similarly “$x, y, z, \ldots$” are the variables in $L_{\text{set}}$ while “$x^\dagger, y^\dagger, z^\dagger, \ldots$” are constants in $L_{\text{set}}$.

Example 0.7. Assume $\rho \in S_{\mathfrak{r}_3}, a \in S_{\mathfrak{c}},$ and $x^\dagger, z^\dagger \in V$. Define the word

$\varphi = \forall \dagger x^\dagger \exists \dagger x^\dagger (\rho(\dagger, x^\dagger, a)) \dagger = (\forall \dagger, x^\dagger, \exists \dagger, z^\dagger, (\dagger, \rho, (\dagger, x^\dagger, z^\dagger, a)^\dagger, (\dagger, \dagger)).$

Words are sets so $\varphi$ is a set. Also $\varphi$ is an $S$-formula i.e. $\varphi \in \Lambda^S$ is a theorem in $T_{\text{set}}$. For simplicity we write

$\varphi = [\forall x \exists z(\rho(x, z, a))]^\dagger.$

Note that $[\ldots]$ and $^\dagger$ (the latter taken by itself) are not symbols in $L_{\text{set}}$ but rather in metalanguage; so, whenever one encounters $[\ldots]^\dagger$ in a text in $L_{\text{set}}$ one needs to replace $[\ldots]^\dagger$ by the corresponding word in $L^*_{\text{set}}$. Now the word $[\exists x \forall \exists z(\rho(x, z, a))]^\dagger$ is not an $S$-formula because constants cannot have quantifiers $\forall, \exists$ in front of them. Also the word $[\forall x \exists z(\rho(x, a))]^\dagger$ is not an $S$-formula because $\rho$ is “supposed to have 3 arguments.” If $\Box \in S_{\mathfrak{f}_2}$ and $x, a, z$ are as above then the word

$\varphi = [\forall x \exists z(\Box (z, a) = z)]^\dagger$

is an $S$-formula.

Remark 0.8. One can define binary operations $\land^*$ and $\lor^*$ on $\Lambda^S$ and a unary operation $\neg^*$ on $\Lambda^S$. For instance, $\land^* : \Lambda^S \times \Lambda^S \rightarrow \Lambda^S$ is defined by

$\varphi \land^* \psi = (\dagger \varphi)^\dagger \land^* (\dagger \psi)^\dagger.$
We have
\[ \varphi \land^* (\psi \land^* \eta) \neq (\varphi \land^* \psi) \land^* \eta, \]
so \( \Lambda_S^* \) is not a Boolean algebra with respect to these operations. For simplicity, and if no confusion arises, we will write \( \land, \lor, \neg \) in place of \( \land^*, \lor^*, \neg^* \).

**Example 0.9.** Let \( M \) be a set. We let \( S_r(M) = M \). For \( n \in \mathbb{N} \) we set \( S_{r_n}(M) = \mathcal{P}(M^n) \), the set of \( n \)-ary relations on \( M \) and \( S_{f_n}(M) \subset \mathcal{P}(M^{n+1}) \) the set of maps \( M^n \to M \). We consider the \( T \)-partitioned set \( S(M) \), union of the above. Then we can consider the formal language \( \Lambda_{S(M)} \). An assignment in \( M \) is a map
\[ \mu : V \to M. \]
For any assignment \( \mu \) there exists a unique map \( v_{M,\mu} : \Lambda_{S(M)}^f \to \{0,1\} \) which is a homomorphism with respect to \( \land, \lor, \neg \), is compatible (in an obvious sense) with \( \forall, \exists \), and satisfies obvious properties with respect to relational and functional symbols, and also with \( \mu \). If \( \varphi \) has no free variables we write \( v_M(\varphi) \) in place of \( v_{M,\mu}(\varphi) \). Here is an example of the “obvious properties” referred to above:
\[ \forall x \forall y \exists y'((x \in M) \land (y \in \mathcal{P}(M))) \to ((v_M(y^t(x)^t) = 1) \leftrightarrow (x \in y))), \]
which translates into argot as follows: for all \( x \in M \) and all \( y \in \mathcal{P}(M) \) the value of \( v_M \) on the sentence
\[ y^t(x)^t = (y, (x,))^t \in \Lambda_S^* \]
is 1 if and only if we have \( x \in y \). Note by the way that
\[ y^t(x)^t \neq [y(x)]^t = y^t(x^t)^t. \]
Here is another example of “obvious property”:
\[ \forall y((y \in M) \to ((v_M(\forall^t x^t (x^t =^t y)) = 1) \leftrightarrow (\forall x((x \in M) \to (x = y))))). \]
Similar properties are satisfied by \( v_M \) for more complicated formulae.

Next we discuss “semantics” of formal languages. The word “semantics” here is taken in a metaphorical sense.

**Definition 0.10.** By a translation of \( \Lambda_S \) into \( \Lambda_{S'} \) we understand a map
\[ m : S \to S' \]
which is compatible with the partitions (in the sense that \( m(S_t) \subset S'_t \) for all \( t \in T \)). For any \( \varphi \in \Lambda_S^* \) one can form, in an obvious way, a formula \( m(\varphi) \in \Lambda_{S'}^f \) obtained from \( \varphi \) by replacing the constants and relational and functional symbols by their images under \( m \), respectively. So we get a map
\[ m : \Lambda_S^* \to \Lambda_{S'}^f \]
which is a homomorphism with respect to \( \land, \lor, \neg \), compatible with \( \forall, \exists \).

An \( S \)-structure (or simply a structure if \( S \) is understood) is a pair \( M = (M, m) \) where \( M \) is a set and \( m \) is a translation
\[ m : S \to S(M). \]
So we get a map
\[ m : \Lambda_S^* \to \Lambda_{S(M)}^f, \]
which is a homomorphism with respect to \( \land, \lor, \neg \), compatible with \( \forall, \exists \). Fix an assignment \( \mu \) and set \( v_{M,\mu} = v_M \). We have a natural map
\[ v_M : \Lambda_{S(M)}^f \to \{0,1\} \]
which is again a homomorphism compatible with $\forall, \exists$. So we may consider the composition

$$v_M = v_M \circ m : \Lambda_f^S \to \{0,1\}.$$ 

We say that a sentence $\varphi \in \Lambda^S$ is satisfied in the structure $M$ if $v_M(\varphi) = 1$. This concept is independent of $\mu$. This is a variant of Tarski's semantic definition of truth: one can define in set theory the predicate is true in $M$ by the definition:

$$\forall x ((x \text{ is true in } M) \leftrightarrow ((x \in \Lambda^S_S) \land (v_M(x) = 1))).$$

(We will continue to NOT use the word true in what follows, though.) If $v_M(\varphi) = 1$ we also say that $M$ is a model of $\varphi$ and we write $M \models \varphi$. (So “is a model” and “$\models$” are predicates that are being added to $L_{set}$.) We say a set $\Phi \subset \Lambda^S_S$ of sentences is satisfied in a structure (write $M \models \Phi$) if all the formulas in $\Phi$ are satisfied in this structure. We say that a sentence $\varphi \in \Lambda^S_S$ is a semantic formal consequence of a set of sentences $\Phi \subset \Lambda^S_S$ (and we write $\Phi \models \varphi$) if $\varphi$ is satisfied in any structure in which $\Phi$ is satisfied. Here the word semantic is used because we are using translations; and the word formal is being used because we are in mathematical logic rather than in pre-mathematical logic. A sentence $\varphi \in \Lambda^S_S$ is valid (or a formal tautology) if it is satisfied in any structure, i.e., if $\emptyset \models \varphi$. We say a sentence $\varphi$ is satisfiable if there is a structure in which it is satisfied. Two sentences $\varphi$ and $\psi$ are semantically formally equivalent if each of them is a semantic formal consequence of the other, i.e $\varphi \models \psi$ and $\psi \models \varphi$; write $\varphi \approx \psi$. Note that the quotient $\Lambda^S_S/\approx$ is a Boolean algebra in a natural way. Moreover each structure $M$ defines a homomorphism $v_M : \Lambda^S_S/\approx \to \{0,1\}$ of Boolean algebras.

**Example 0.11.** Algebraic structures can be viewed as models. Here is an example. Let

$$S = \{\star, \iota, e\}$$

with $e$ a constant symbol, $\star$ a binary function symbol, and $\iota$ a unary function symbol. Let $\Phi_{gr}$ be the set of $S$-formulas $\Phi_{gr} = \{\varphi_1, \varphi_2, \varphi_3\}$ where

$$\begin{align*}
\varphi_1 &= \forall x \forall y \forall z (x \star (y \star z) = (x \star y) \star z) \\
\varphi_2 &= \forall x (x \star e = e \star x = x) \\
\varphi_3 &= \forall x (x \star \iota(x) = \iota(x) \star x = x).
\end{align*}$$

Then a group is simply a model of $\Phi_{gr}$ above. We also say that $\Phi_{gr}$ is a set of axioms for the formalized theory of groups.

More generally:

**Definition 0.12.** A formalized (or formal) system of axioms is a pair $(S, \Phi)$ where $S$ is a $T$-partitioned set and $\Phi$ is a subset of $\Lambda^S_S$; then define

$$\Theta = \Phi^= = \{\varphi \in \Lambda^S_S; \emptyset \models \varphi\};$$

$\Theta$ is called the formal theory generated by the system of axioms $(S, \Phi)$.

**Remark 0.13.** One thinks of $\Phi^=\$ as semantically defined because its definition involves models and hence translations. One may define a “syntactic” version of this set namely the set

$$\Phi^+ = \{\varphi \in \Lambda^S_S; \emptyset \vdash \varphi\};$$

where $\vdash$ is the predicate added to $L_{set}$ meaning “$\varphi$ provable from $\Phi$” in an obvious sense that imitates the definition of proof in pre-mathematical logic. (For this to work one needs a formal analogue of a witness assignment; the latter is defined
as a function with the obvious properties. For instance one can replace $\Lambda_S$ by an obviously defined set that plays the role of “witness closure”.) We will not make this precise here, neither will we use the following theorem of Gödel which is usually referred to as Gödel’s completeness theorem. This theorem intuitively says that syntactic and semantic provability coincide.

**Theorem 0.14.** $\Phi^c = \Phi^=.$

**Exercise 0.15.** Write down a set of symbols $S$ and a set of formulas $\Phi$ which is a set of axioms for the formalized theory of:
1) commutative unital rings;
2) fields;
3) ordered sets;
4) vector spaces over a given field;
5) Boolean algebras;
6) sets.
Can one find such a pair $S, \Phi$ in the following cases?
7) well ordered sets;
8) topological spaces;
9) the ring of integers.
Chapter 48: Incompleteness

(alternative version)

Our aim here is to state and explain some classical incompleteness theorems due to Gödel and Cohen. Recall that all definitions in our present context are definitions in the language \( L \) and all metadefinitions are in our fixed Metalanguage. Assume \( L \) has constants \( c_1, c_2, \ldots \), e.g. assume \( L \) has only witnesses as constants and we ordered the witnesses in a “sequence”.

**Definition 0.16.** Define the sets \( \in^\dagger, c_1^\dagger, c_2^\dagger, \ldots, S_\text{set}, \Lambda_\text{set} \) by

\[
\in^\dagger = \{77\} \quad \text{(say)},
\]

\[
c_1^\dagger = \{c_1\},
\]

\[
c_2^\dagger = \{c_2\},
\]

....

\[
S_\text{set} = \{\in^\dagger, c_1^\dagger, c_2^\dagger, \ldots\},
\]

\[
\Lambda_\text{set} = \Lambda_{S_\text{set}}.
\]

**Example 0.17.** Assume, for instance, that \( c_5 = c_{x=x} \). Then

\[
c_5^\dagger = \{c_{x=x}\} = c^{\exists x((z \in y) \leftrightarrow (z = c_{x=x}))}.
\]

**Metadefinition 0.18.** For any formula \( P \) in \( L^f_\text{set} \) denote by \( P^\dagger \) the formula in \( \Lambda^f_\text{set} \) obtained from \( P \) by replacing all the symbols with the same symbols carrying the superscript \( \dagger \).

So “\( P^\dagger \in \Lambda^f_\text{set} \)” is a theorem in \( T_\text{set} \) (that is proved directly from the definitions). If \( P \) is a sentence in \( L^f_\text{set} \) then \( P^\dagger \) is a sentence in \( \Lambda^f_\text{set} \).

**Example 0.19.** For instance if \( P \) is the sentence “\( \exists x (x \in b) \)” in \( L^f_\text{set} \) then

\[
P^\dagger = [\exists x (x \in b)]^\dagger = \exists^\dagger x \{^\dagger x \in ^\dagger b\}^\dagger = (\exists^\dagger x, ^\dagger x, (^\dagger x, ^\dagger, b^\dagger), ^\dagger) \in \Lambda^f_\text{set}.
\]

Note that \( P^\dagger \) is a word, hence a set, hence a constant in \( L_\text{set} \). So, using appropriate definitions, we attached to a string \( P \) consisting of a quantifier “\( \exists \)” a variable “\( x \)” a separator “(,” etc., in \( L_\text{set} \), a single constant “\( P^\dagger \)” in \( L_\text{set} \). As before, \([.,]^\dagger \) are not symbols in \( L_\text{set} \) but rather in metalanguage; so, whenever one encounters \( P^\dagger \) or \([.,]^\dagger \) in a text in \( L_\text{set} \) one needs to replace \( P^\dagger \) (or \([.,]^\dagger \) respectively) with the corresponding word in \( L^*_\text{set} \).

Let \( B_1, B_2, \ldots \) be the ZFC axioms in \( L^*_\text{set} \).

**Definition 0.20.** Formalized (or formal) set theory is the formal theory \( \Theta_\text{set} \) generated by \( (S_\text{set}, \Phi_\text{set}) \) where \( \Phi_\text{set} = \{B_1^\dagger, B_2^\dagger, \ldots\} \).
Remark 0.21. Note that $L_{\text{set}}$ is a text (more precisely a collection of symbols) whereas $\Lambda_{\text{set}}$ is a set (hence, again, a text, but consisting of one symbol only) hence $\Lambda_{\text{set}}$ is a constant in $L_{\text{set}}$. Similarly the theory $T_{\text{set}}$ is a text (more precisely a collection of strings of symbols) while the formal theory $\Theta_{\text{set}}$ is itself a set (hence, again, a text consisting of one symbol only), hence a constant in $L_{\text{set}}$.

Definition 0.22. A formal theory $\Theta$ generated by a system of axioms $(S, \Phi)$ is called formally inconsistent if there exists a sentence $\varphi \in \Lambda_{\text{set}}$ such that $\varphi \in \Theta$ and $\neg \varphi \in \Theta$. $\Theta$ is called formally consistent if it is not formally inconsistent. $\Theta$ is called formally complete if for any sentence $\varphi \in \Lambda_{\text{set}}$ either $\varphi \in \Theta$ or $\neg \varphi \in \Theta$. $\Theta$ is called formally incomplete if it is not formally complete.

An immediate consequence of Theorem 0.14 is the following theorem in $T_{\text{set}}$:

Theorem 0.23. A formal theory $\Theta$ generated by a system of axioms $(S, \Phi)$ is formally consistent if and only if $\Phi$ has a model.

Let $P_{\text{con}}$ be the sentence in $L_{\text{set}}$ (in set theory) expressing the formal consistency of formal set theory $\Theta_{\text{set}}$; in other words

$$P_{\text{con}} \text{ equals } "\Theta_{\text{set}} \text{ is formally consistent}"$$

equivalently

$$P_{\text{con}} \text{ equals } "\neg(\exists x((x \in \Theta_{\text{set}}) \land ((\neg x) \in \Theta_{\text{set}})))".$$  

Recall that $\neg = \neg^\dagger : \Lambda_{\text{set}} \to \Lambda_{\text{set}}$ is the function $\neg(\varphi) = \neg^\dagger(\dagger \varphi)^\dagger$.

We have the following trivial theorem in $T_{\text{set}}$:

Theorem 0.24. $P_{\text{con}} \leftrightarrow (\Theta_{\text{set}} \neq \Lambda_{\text{set}})$.

Consider $P_{\text{con}}^\dagger$ as a set belonging to $\Lambda_{\text{set}}$. Then Gödel proved the following deep theorem in $T_{\text{set}}$:

Theorem 0.25. $P_{\text{con}} \rightarrow (P_{\text{con}}^\dagger \notin \Theta_{\text{set}})$.

A translation of the above into English would be: if set theory is consistent then its consistency cannot be proved within set theory. The latter is, of course, horrifyingly imprecise.

Remark 0.26. Note Theorem 0.25 says nothing about the consistency of ZFC. And indeed Theorem 0.25 would be a useless statement if ZFC itself were not consistent.

With regards to the continuum hypothesis we have the following results due to Gödel and Cohen, respectively. Let $P_{\text{CH}}$ be the sentence in $L_{\text{set}}$ expressing the continuum hypothesis, i.e. expressing that any subset of $\mathcal{P}(\mathbb{N})$ is either in bijection to $\mathbb{N}$ or in bijection to $\mathcal{P}(\mathbb{N})$. We have $P_{\text{CH}}^\dagger \in \Lambda_{\text{set}}$ is a theorem in $T_{\text{set}}$. Then we have the following theorems in $T_{\text{set}}$:

Theorem 0.27. $P_{\text{con}} \rightarrow (\neg P_{\text{CH}}^\dagger \notin \Theta_{\text{set}})$.

Theorem 0.28. $P_{\text{con}} \rightarrow (P_{\text{CH}}^\dagger \notin \Theta_{\text{set}})$.

Note that Theorems 0.27 and 0.28 imply:

Corollary 0.29. If $\Theta_{\text{set}}$ is formally consistent then $\Theta_{\text{set}}$ is formally incomplete.
The principle of proof of Cohen’s Theorem 0.28 is simple: in view of Theorem 0.14 it is enough to find a model of $\Phi_{\text{set}}$ which is not a model for $\Phi_{\text{set}} \cup \{P_{\text{CH}}^\dagger\}$. Proving that such a model exists is hard.

The principle of proof of Gödel’s Theorem 0.27 is less obvious. We sketch below the rough idea of the proof following (Cohen 2008). Gödel’s main idea was to define a unary predicate in $L_{\text{set}}$ that he called $L(x)$ whose translation in English is “$x$ is constructible”; intuitively (and very vaguely) $x$ is constructible if it can be defined by “transfinite induction” using formulae in ZFC. Then he defined a function

$$(L : \Lambda_{\text{set}}^L \rightarrow \Lambda_{\text{set}}^L, \varphi \rightarrow \varphi_L)$$

sending any formula $\varphi$ into the formula $\varphi_L$ obtained from $\varphi$ by “asking that all bound variables satisfy $L$”; e.g. if $\varphi$ is

$$[\forall x((x \in y) \rightarrow (x \in z))]^\dagger$$

then $\varphi_L$ is

$$[\forall x((L(x) \rightarrow ((x \in y) \rightarrow (x \in z))))]^\dagger.$$

(A similar device was used in the discussion of metamodels.) Next one proves the following theorems in $T_{\text{set}}'$. Using Gödel’s notation we write $A = \forall xL(x)$.

**Theorem 0.30.** $(\Theta_{\text{set}})_L \subset \Theta_{\text{set}}$.

**Theorem 0.31.** $A^L_L \in \Theta_{\text{set}}$.

**Theorem 0.32.** $[A \rightarrow P_{\text{CH}}]^\dagger \in \Theta_{\text{set}}$.

Now note that Theorems 0.30, 0.31, 0.32 above imply Theorem 0.27. To see this assume $P_{\text{con}}$ (i.e. $\Theta_{\text{set}}$ is formally consistent), assume

$$\neg P_{\text{CH}}^\dagger \in \Theta_{\text{set}},$$

and seek a contradiction. By Theorem 0.30 we get

$$(\neg P_{\text{CH}}^\dagger)_L \in \Theta_{\text{set}}.$$

Since $[ ]^L_L$ commutes with $\neg$ we get:

(0.3) $$\neg ((P_{\text{CH}}^\dagger)_L) \in \Theta_{\text{set}}.$$  

By Theorems 0.30 and 0.32 we get

$$[A \rightarrow P_{\text{CH}}]^\dagger \in \Theta_{\text{set}}.$$  

Now since $[ ]^L_L$ commutes with the connectives we get

$$(A^L_L \rightarrow (P_{\text{CH}}^\dagger)_L) \in \Theta_{\text{set}}.$$  

By Theorem 0.31 and modus ponens in $\Theta_{\text{set}}$ we get

(0.4) $$(P_{\text{CH}}^\dagger)_L \in \Theta_{\text{set}}.$$  

Now 0.3 and 0.4 imply that

$$(P_{\text{CH}}^\dagger)_L \land (\neg (P_{\text{CH}}^\dagger)_L) \in \Theta_{\text{set}}$$

hence $\Theta_{\text{set}}$ is not formally consistent, a contradiction; so Theorem 0.27 follows.

Next we want to analyze the integers in mathematical logic.
Remark 0.33. The Peano axiom in set theory is the sentence that there exists a Peano triple. One can view this as a sentence in $\Lambda_{\text{set}}$ of formalized set theory; indeed the words “there exists a set $z$ such that for any subset $y$ ...” can be written as

$$\exists z \forall y ((\forall x (x \in y \rightarrow (x \in z)) \rightarrow ...).$$

But if one tries to write this axiom, say, in the formal language

$$\Lambda_{\text{ar}} = \Lambda_{\{+,x,1,s,0,1,2,...\}}$$

where $+,x$ are binary functions, $s$ is a unary function, and $0,1,2,...$ are constants, then one is bound to fail because there is no way to say, in this formal language, that for any subset something happens. (If one says $\forall x$ then in any $S$-structure $(M,m)$ the variable $x$ is translated as an element of $M$ and not as a subset of $M$.) Nevertheless there is a formal Peano system of axioms in the formal language $\Lambda_{\text{ar}}$ which we are going to consider below.

Definition 0.34. Consider the set $S = S_{\text{ar}} = \{+,x,1,s,0,1,2,...\}$ where $+,x$ are binary functional symbols, $s$ is a unary functional symbol, and $0,1,2,...$ are constants. Let $\Phi_{\text{ar}}$ be the following sentences in $\Lambda_{\text{ar}} = \Lambda_{S_{\text{ar}}}$ (called the formal Peano axioms): $A_1, A_2, A_3$ (obtained from the Peano axioms). So this is a countable set of axioms; the latter series of axioms stands for a weak form of induction. The formal system of axioms of Peano arithmetic is the pair $(S_{\text{ar}}, \Phi_{\text{ar}})$. The theory $\Theta_{\text{ar}}$ generated by $(S_{\text{ar}}, \Phi_{\text{ar}})$ is called the Peano arithmetic.

We have the following trivial theorem in $T_{\text{set}}$:

Theorem 0.35. $\Theta_{\text{ar}}$ is formally consistent.

Proof. Indeed $\Theta_{\text{ar}}$ has a model $N = (\mathbb{N}, n)$, where $n : S_{\text{ar}} \rightarrow S(\mathbb{N})$ sends $i \mapsto i$ for $i \in \mathbb{N}$, and sends $n(+^1) = +^N$, where $+^N \subset \mathbb{N} \times \mathbb{N} \times \mathbb{N}$ is the (graph of) the addition, etc. $\square$

Remark 0.36. Note that the formal consistency of $\Theta_{\text{ar}}$ in Theorem 0.35 says nothing about the consistency of the Peano axioms.

The model $N = (\mathbb{N}, n)$ in the proof of Theorem 0.35 is called the standard model of arithmetic. Let $P$ be any formula in $L_{\text{set}}'$ such that $P^1 \in A_{\text{ar}}$. (Here we assume $L_{\text{set}}$ is enriched to include $(\mathbb{N}, +, \times)$.) Also let $P^N$ be the formula in $L_{\text{set}}'$ obtained from $P$ by inserting the condition that all variables and constants belong to $\mathbb{N}$. Then, by definition of $v_N$, if $P$ is a sentence,

$$(v_N(P^1) = 1) \iff P^N$$

is a theorem in $T_{\text{set}}$. Note then that, in $T_{\text{set}}$, we have:

Theorem 0.37. $(P^1 \in \Theta_{\text{ar}}) \rightarrow P^N$.

In English this could be translated as: If there is a proof that $P$ in the language of arithmetic then $P$, interpreted as a statement about $\mathbb{N}$, is a theorem in set theory. The latter formulation is, of course, very imprecise.

Proof. Assume $P^1 \in \Theta_{\text{ar}}$ and let $N = (\mathbb{N}, n)$ be the standard model of arithmetic. Then, by the definition of $\Theta_{\text{ar}}$, we have $v_N(P^1) = 1$ for any model $M$ of $\Theta_{\text{ar}}$, in particular $v_N(P^1) = 1$; hence, by 0.5, we have $P^N$ which ends the proof. $\square$
Remark 0.38. The above argument breaks down if instead of $\Theta_{ar}$ we take $\Theta_{set}$ because, in this case, we do not have an analogue (appropriately well behaved syntactically) of the standard model $\mathbb{N} = (\mathbb{N}, n)$.

Here is Gödel’s incompleteness theorem of arithmetic. It is a theorem in $T_{set}$.

Theorem 0.39. $\Theta_{ar}$ is formally incomplete.

Sketch of proof. We follow the informal presentation in the first pages of Gödel’s original article, in combination with facts proved later in that paper. We refer to loc. cit. for the details. All of this is adapted to our paradigm that distinguishes carefully between sentences $P$ and their images $P^\dagger$.

The first step is to “tag” formulae by integers. Choose an explicit bijection between $\Lambda_{ar}$ and $\mathbb{N}$ and write

$$\Lambda_{ar} = \{s_1, s_2, s_3, \ldots\}.$$  

Let $p : \mathbb{N} \to \mathbb{N}$ be the function defined by letting $p(i)$ be the $i$th prime. Then define the “Gödel numbering function”

$$G : \Lambda_{ar}^* \to \mathbb{N}$$

by the rule

$$G(s_1 s_2 s_3 \ldots) = p(1)^{i_1} p(2)^{i_2} p(3)^{i_3} \ldots = 2^{i_1} 3^{i_2} 5^{i_3} \ldots$$

Clearly $G$ is injective and “definable in $L_{ar}$” in the sense that its graph is given by an appropriate formula in $L_{ar}^f$ in two variables (where, again, we identify $\Lambda_{ar}$ with $\mathbb{N}$ in some explicit way). Let $\Sigma = G(\Lambda_{ar}^f)$ and $\Sigma'$ be the image by $G$ of the set of all formulae with exactly one free variable. Let $F : \Sigma \to \Lambda_{ar}^f$ be the inverse of $G : \Lambda_{ar}^f \to \Sigma$.

The second step is to show that the predicate “belongs to $\Theta_{ar}$” is “definable” in $L_{ar}$ in the following sense. One (explicitly!) defines a predicate $B$ by a formula in $L_{ar}^f$ with one free variable and one proves (by induction) that for all $n \in \Sigma$,

$$(0.6) \quad (F(n) \in \Theta_{ar}) \iff B(n).$$

This is the heart of the proof: it amounts to showing that “provability of a sentence is an arithmetic property of the tag of the sentence.” This is very laborious but not exceedingly hard. (What one uses is the syntactic description of provability: $\Theta_{ar} = \Phi_{ar}^\dagger$; cf Theorem 0.14. As a baby example of what is being done say that $\forall^\dagger = s_6$; then syntactic conditions such as “$\varphi$ does not contain $\forall$” are expressed in the language of arithmetic as “if $p$ is a prime and $p^6 | G(\varphi)$ then $p^7 | G(\varphi)$.”)

The third step is to (explicitly!) define a predicate $S$ by a formula in $L_{ar}^f$ in 3 variables “defining” the following “substitution” function $S^f : \Sigma' \times \mathbb{N} \to \mathbb{N}$: $S^f(a, b) = c$ if and only if $F(c)$ is the sentence obtained from the formula $F(a)$ by replacing the unique free variable of $F(a)$ by $b$; in other words

$$(0.7) \quad F(S^f(a, b)) = F(a)(b).$$

Now there is a formula $P$ in $L_{set}^*$ such that

$$(0.8) \quad \forall x (P(x) \leftrightarrow ((x \in \Sigma) \land (\neg B(S(x, x)))))$$

and $P(x)^\dagger$ is in $\Lambda_{ar}^f$. Since the latter has exactly one free variable $x^\dagger$ there exists $m \in \Sigma$ such that

$$(0.9) \quad P(x)^\dagger = F(m).$$
Now consider the sentence \( P^{\text{Sn}}(m) \) in \( L^*_\text{set} \). By 0.8,
\[
(0.10) \quad P^{\text{Sn}}(m) \leftrightarrow (\neg(B(S^\dagger(m,m)))).
\]
On the other hand, by 0.7 and 0.9,
\[
(0.11) \quad F(S^\dagger(m,m)) = F(m)(m) = P(m)^\dagger,
\]
hence \( P(m)^\dagger \) is in \( \Lambda^*_\text{ar} \) and
\[
(0.12) \quad S^\dagger(m,m) = G(P(m)^\dagger).
\]
Hence
\[
(0.13) \quad \neg P^{\text{Sn}}(m) \leftrightarrow \neg(B(S^\dagger(m,m))) \quad \text{by 0.10}
\]
\[
\leftrightarrow B(S^\dagger(m,m)) \quad \text{by 0.12}
\]
\[
\leftrightarrow (P(m)^\dagger \in \Theta_{\text{ar}}) \quad \text{by 0.6}
\]
\[
\rightarrow P^{\text{Sn}}(m) \quad \text{by Theorem 0.37}.
\]
The last step of the proof is to check the following:
\[
(0.14) \quad P(m)^\dagger \notin \Theta_{\text{ar}};
\]
\[
(0.15) \quad \neg P(m)^\dagger \notin \Theta_{\text{ar}}.
\]
To check 0.14 assume \( P^\dagger(m) \in \Theta_{\text{ar}} \) and seek a contradiction. By 0.13 we get both \( P^{\text{Sn}}(m) \) and \( \neg P^{\text{Sn}}(m) \) which is a contradiction.
To check 0.15 assume \( \neg P(m)^\dagger \in \Theta_{\text{ar}} \) and seek a contradiction. Since by Theorem 0.37
\[
(\neg P(m)^\dagger \in \Theta_{\text{ar}}) \rightarrow (\neg P^{\text{Sn}}(m))
\]
we get \( \neg P^{\text{Sn}}(m) \). By 0.13 we get \( P^{\text{Sn}}(m) \). This is a contradiction and our Theorem is proved. \( \square \)

**Remark 0.40.** Since Theorems 0.25, 0.27, 0.28, 0.39 are sentences in \( L^*_\text{set} \), they have, according to the conventions in the present course, no meaning and no truth value. In particular these theorems say nothing about mathematics itself. One could ask, however, what happens if we declare at this point that \( L_{\text{set}} \) actually has a reference (which we ignored so far but we are ready now to take into consideration) and that \( L_{\text{set}} \) is not self-referential (which we have to be careful to insist on); is it the case, under these new assumptions, that Theorems 0.25, 0.27, 0.28, 0.39 do actually say something about mathematics itself? We will give an argument below that, even under these new assumptions, these theorems still cannot be seen (in our paradigm) as saying anything about mathematics itself. Indeed let us take Theorem 0.25 as an example. It can be reformulated as the following sentence in \( L_{\text{set}} \):
\[
(0.16) \quad (\Theta_{\text{set}} \neq \Lambda^*_{\text{set}}) \rightarrow ((\Theta^\dagger_{\text{set}}), \neq \dagger, \Lambda^*_{\text{set}}) \notin \Theta_{\text{set}}).
\]
A statement about mathematics itself (i.e. about \( T_{\text{set}}, L_{\text{set}}, \) etc.) which would be of interest and that seems to correspond to 0.16 would be the following metasentence in the metalanguage \( \tilde{L}_{\text{set}} \) (that talks about \( L_{\text{set}} \)):
\[
(0.17) \quad \text{If } T_{\text{set}} \neq L_{\text{set}} \text{ then "} T_{\text{set}} \neq L_{\text{set}} \text{" is not in } T_{\text{set}}.
\]
Now all we seem to have to do is try to consider 0.17 as a translation of 0.16. However this is fallacious: indeed, under our assumptions, there is no translation from $L_{set}$ into $\hat{L}_{set}$ (as one can easily see from the fact that we assumed $L_{set}$ has a reference but is not self-referential and translations preserve reference.) We reached the conclusion that,

Even if we allow the language $L_{set}$ of mathematics to have a (non-linguistic) reference, Gödel’s theorems (and indeed mathematical logic as a whole) cannot say anything about mathematics itself.

Of course this conclusion was a corollary of our formalist (nominalist) position. On the other hand Gödel himself was a Platonist (metaphysical realist) and his own conclusions with regards to what his theorems are saying about mathematics were rather different from the one expressed in this course. For more on Gödel’s Platonist views see (Cohen 2008) and (Wang 1996).

Remark 0.41. Let $T'_{set}$ be the theory generated by $T_{set}$ and all sentences of the form

$$(P^1 \in \Theta_{set}) \rightarrow P$$

where $P$ is a sentence in $L'_{set}$. Intuitively $T'_{set}$ postulates that “provability of formal provability implies provability.” Similarly one can consider the theory $T''_{set}$ generated by $T_{set}$ and all sentences of the form

$$P \rightarrow (P^1 \in \Theta_{set}).$$

Intuitively $T''_{set}$ postulates that “provability implies provability of formal provability.” Finally one can consider the theory $T'''_{set}$ generated by $T_{set}$ and all sentences of the form

$$P \leftrightarrow (P^1 \in \Theta_{set}).$$

Intuitively $T'''_{set}$ postulates that “provability of formal provability is equivalent to provability.” We would like to make some comments on the theories $T'_{set}, T''_{set}, T'''_{set}$. Clearly $T'_{set}$ and $T''_{set}$ are both contained in $T'''_{set}$.

Note now that $P_{con}$ is a theorem in $T'_{set}$. Indeed, assume, in $T'_{set}$, that $\neg P_{con}$ and seek a contradiction. Then $\Theta_{set} = \Lambda_{set}'$, hence $[1 \neq 1]^1 \in \Theta_{set}$ hence $1 \neq 1$, a contradiction.

Note also that $\neg P_{con}$ is a theorem in $T''_{set}$. Indeed, assume, in $T''_{set}$, that $P_{con}$, and seek a contradiction. Then, by Theorem 0.25 $P_{con}^1 \notin \Theta_{set}$. So, by the “new” axioms of $T'''_{set}$, we get $\neg P_{con}$, a contradiction.

By the above we get that $P_{con} \land \neg P_{con}$ is a theorem in $T'''_{set}$ so $T'''_{set}$ is inconsistent. It is not clear a priori if $T'_{set}$ or $T''_{set}$ are inconsistent. The theory $T'_{set}$ seems to be, morally, self-defeating because $\neg P_{con}$ is a theorem in this theory (which is definitely something that we do not want, although it is not something that implies inconsistency). On the other hand $T'_{set}$, in which $P_{con}$ is a theorem, seems to be a reasonable extension of $T_{set}$ (unless found to be inconsistent).

Exercise 0.42. Show that in all arguments given in Remark 0.41, Theorem 0.25 can be replaced by Theorems 0.27 and 0.28.

Exercise 0.43. Here is an attempt to prove that $(P^1 \in \Theta_{set}) \leftrightarrow P$ in $T_{set}$. This attempt is wrong; explain why it is wrong. (If the attempt were correct, by the way, we would have, due to Remark 0.41, a rather convincing metaproof that
\( T_{set} \) is inconsistent which would be the end of ZFC.) Say we want to prove the implication \( \leftarrow \). Assume \( P \). Then there exists a proof \( A, B, C, \ldots, P \) of \( P \) in \( T \). We may then consider the sequence \( A^\dagger, B^\dagger, C^\dagger, \ldots, P^\dagger \) which is a formal proof of \( P^\dagger \); this shows that \( P^\dagger \in \Phi_{set} \models = \Theta_{set} \). A similar argument could be given for the implication \( \rightarrow \). Hint: the mistake is that the above is not a proof in \( L_{set} \) but a metaproof in the metalanguage \( \hat{L} \); analyze the situation in more detail.
Bibliography