The Use of Trustworthy Principles in a Revised Hilbert's Program

Anton Setzer*

November 11, 2015

Abstract

After the failure of Hilbert's original program due to Gödel's second incompleteness theorem, relativized Hilbert's programs have been suggested. While most metamathematical investigations are focused on carrying out mathematical reductions, we claim that in order to give a full substitute for Hilbert's program, one should not stop with purely mathematical investigations, but give an answer to the question why one should believe that all theorems proved in certain mathematical theories are valid.

We suggest that, while it is not possible to obtain absolute certainty, it is possible to develop trustworthy core principles using which one can prove the correctness of mathematical theories. Trust can be established by both providing a direct validation of such principles, which is necessarily non-mathematical and philosophical in nature, and at the same time testing those principles using metamathematical investigations. We investigate three approaches for trustworthy principles, namely ordinal notation systems built from below, Martin-Löf type theory, and Feferman's system of explicit mathematics. We will review what is known about the strength up to which direct validation can be provided.

1 Reducing Theories to Trustworthy Principles

In the early 1920's Hilbert suggested a program for the foundation of mathematics, which is now called Hilbert's program. As formulated in [40], "it calls for a formalization of all of mathematics in axiomatic form, together with a proof that this axiomatization of mathematics is consistent. The consistency proof itself was to be carried out using only what Hilbert called 'finitary' methods. The special epistemological character of finitary reasoning then yields the required justification of classical mathematics." Because of Gödel's second incompleteness theorem, Hilbert's program can be carried out only for very weak theories. Because of this failure (see, e.g., [44, 40]) a relativized Hilbert's program has

*Department of Computer Science, Swansea University, Singleton Park, Swansea SA2 8PP, UK, Email: a.g.setzer@swan.ac.uk, http://www.cs.swan.ac.uk/∼csetzer/, Tel: +44 1792 513368, Fax: +44 1792 295651.

been suggested by Kreisel (Zach [44] cites [17, 18, 19]), and then further developed by Feferman ([7, 8, 9, 10]). In the approach by Feferman [7, 9], one considers two frameworks $\mathcal{F}_1$ and $\mathcal{F}_2$. $\mathcal{F}_1$ could mean infinitary, $\mathcal{F}_2$ finitary, or $\mathcal{F}_1$ mean nonconstructive, $\mathcal{F}_2$ constructive (see p. 367 of [7]). Consider for $i \in \{1, 2\}$ certain theories T_i formulated in languages $\mathcal{L}_i$ corresponding to frameworks $\mathcal{F}_i$. Let Φ be a primitive recursive subset of the formulae of $\mathcal{L}_1 \cap \mathcal{L}_2$. Let U be a third theory, usually a very weak theory such as PRA. Then combining [8, 10], we have $T_1 \leq T_2[\Phi]$ in U, if there exists a partial recursive function f such that

(1) if p is a proof in T_1 of a formula φ in Φ, then $f(p)$ is a proof of φ in T_2;

(2) (1) can be shown in U.

Feferman presents many examples of such reductions.

This program of reductive proof theory gives rise to many interesting connections between various theories which provides us with a broad picture of mathematical theories and their relationship. While being very insightful and resulting in lots of metatheorems, it fails to answer the initial question by Hilbert, namely: do I know that my original theory T_1 is consistent? Or widening it in the sense of Kreisel and Feferman: If I have proved in theory T_1 a mathematical statement, do I know that it is valid? If we take say a proof of Fermat's last theorem, do we know that there is actually no counter example to this theorem? From Gödel's second incompleteness theorem it follows that there is no mathematical argument that excludes that there is at the same time a proof of Fermat's last theorem in a theory T_1 and a counter example (unless T_1 is very weak), without assuming at least the consistency of another theory of at least equal strength.

Many mathematicians evade this problem and say that all they want is to have a proof which can be formulated in, for instance, Zermelo-Fraenkel set theory. However, this is not what mathematics is intended for. Mathematics is not just a glass bead game in the sense of Hesse [15], a formal game of finding strings of symbols which follow certain decidable rules. The goal of mathematics is, as any science, to establish truth about real properties. In case of Fermat's last theorem, we want to know whether there are no numbers violating it.

What we can do, in the sense of Kreisel and Feferman, is to reduce T_1 to another theory T_2, which is essentially as strong as T_1, and then obtain that T_2 proves as well the mathematical theorems of T_1 we are interested in. Any mathematical argument will only reduce T_2 further to another theory T_3. So in order not to continue going in circles, we need to reduce T_1 to one theory T_2 for which we can give reasons why we believe that everything it proves (possibly restricted to a subset of statements) is valid.

At this point pure mathematical reasoning ends. No matter what we do, we cannot obtain absolute certainty. However we can establish trust. Trust does not mean blind faith. Trust is established by convincing ourselves in the best possible way that what we trust in does not break. This means that we carefully investigate the principles underlying T_2, examine them, and give an

argument why we can trust them. However, such an analysis can never be done in a purely mathematical way – if we do this, then we just reduce T_2 to a third theory T_3, namely the theory in which the argument of the correctness of T_2 is formulated, and we just have added a new theory to our chain of theories.

However, what we can do, and many constructive and semi-constructive theories have been developed for this purpose, is to formulate theories T_2 where these principles are as pure and clean as possible. Then we can carry out two further steps:

(1) We can formulate as precisely as possible an argument why we believe that we can trust in those principles. Note that this is no longer a purely mathematical argument. However, making it as precise as possible is a very valuable exercise, since it could reveal any possible flaws in those principles.

(2) Since an argument as in (1) does not have the status of a mathematical theorem, it can never provide absolute certainty.[1] Therefore what is needed is to carry out additional testing. Note that mathematicians will in many cases still test their mathematical theorems even if they have proven them, however usually only in order to detect possible flaws in their proofs.

How do we test a theory?

- We can look at theorems provable in T_2 and check whether the theorems actually are true (e.g. in case of Fermat's last theorem that there is no counter example). However, there is one problem, namely that by the results of reverse mathematics we know that most mathematical theorems require very little proof theoretic strength. So such tests do not explore the limits of the theory.

  Peter Dybjer has in [3] suggested to develop meaning explanations for Martin-Löf type theory based on the principle that for each judgement of type theory a test is given. The judgement is valid if it passes all tests. Once carried out in full ([3] provides only the basic idea) one obtains for every provable judgements of type theory a test for its validity. Dybjer's article was a major inspiration for this part of the article.

- Ordinal analysis, or any other proof theoretic analysis (e.g. normalisation proofs) is a very strong test, because it tests the theory at its limits. However, this does not establish absolute certainty. When the author was pointing out to Per Martin-Löf that Michael Rathjen had told the author that he knows that Π^1_2-CA is consistent because he has proof theoretically analysed it, Martin-Löf pointed out that he had an inconsistent type theory and a normalisation proof of it. The problem was that the normalisation proof was carried out in an inconsistent theory. So even a cut elimination or normalisation argument does not guarantee the consistency of the theory.

[1] Of course even mathematical theorems can never give absolute certainty as outlined before. One can think as suggested by one of the referees that a short carefully checked mathematical proof that uses no controversial principles is the paradigm of practical certainty. However, unless one uses extremely weak principles, Gödel's incompleteness theorem applies here as well – even though it is unlikely that an inconsistency is used, we cannot exclude it.

Does this mean that we should give up proof theoretical analysis and normalisation proofs? No, not at all. If a theory is inconsistent, it is likely but not guaranteed that the inconsistency will be found when analysing it proof theoretically. A proof theoretic analysis is up to now one of the strongest ways to stretch a theory to its limits, because it requires to use principles which cannot be reduced to simpler ones. We can often reduce theories which are more expressive to less expressive ones of equal strength in such a way that the reduction shows that they are equiconsistent. However, we cannot reduce a proof theoretic stronger theory to a weaker one, unless both are inconsistent. A proof theoretic analysis needs to distinguish theories of different strength and therefore needs to make use of the principles which are responsible for its strength and which cannot be reduced to weaker ones.

One reason why a proof theoretic analysis is of big significance was pointed out by one of the referees of this article, who wrote "Something that makes specifically ordinal-theoretical proof-theoretical analyses of a theory particularly convincing is that in many cases there is a big difference between the metatheory and the object theory; whereas with normalisation proofs based on Tait-style computability, or Girard-style 'candidates', the metatheory is (more-or-less) the theory itself together with a uniform reflection principle. Something would be far wrong if one could not prove a normalisation theorem for Church's theory of types in such a metatheory; but the extra confidence one gets in the principles formulated therein from a normalisation theorem is tiny."

- In general, any metamathematical analysis of a theory is a test of it. It requires to investigate all axioms and rules of the theory in detail. And if there is an inconsistency in a theory, there is the possibility that one discovers it when carrying out this analysis.[2] If one does not discover any problem, we know at least that any derivation of an inconsistency must be increasingly complicated, since it escaped such a careful analysis. So even if a theory is eventually found to be inconsistent, it is likely that most proofs carried out in it do not make use of it, and we can replace them by proofs in a weaker theory, which does not have this inconsistency.

Therefore there is the need to define mathematical theories in which we can put our trust and describe as clearly as possible the reasons why we trust in the consistency of those theories.

1.1 Does the Consistency Problem Matter?

When discussing the problem about consistency, many mathematicians will wonder why there is a problem. Zermelo-Fraenkel set theory (ZF) has been in use

[2]However, we can never be certain since the metatheory in which the analysis is carried out would be inconsistent as well.

since 1922. Most of mathematics can be carried out in extensions of it, and it has been analysed thoroughly by set theorists.

However, as we know from reverse mathematics, most mathematical proofs can be carried out in theories which are proof theoretically very weak compared to ZF, therefore mathematical proofs will not explore the limits of ZF. Metamathematical investigations have not really stretched theories having the strength of ZF or greater by themselves, but only investigated such theories relative to other theories of strength of at least that of ZF. Proof theory has succeeded to analyse in unpublished form (Arai, [1], see as well [2]) theories of strength Kripke-Platek set theory + Π_1-Collection + V = L (which embeds $(\Sigma^1_3 - \mathrm{DC}) + \mathrm{BI}$ and $(\Sigma^1_3 - \mathrm{AC}) + \mathrm{BI}$). In fully published form Rathjen has analysed [33] the theory of Kripke-Platek set theory plus the existence of one stable ordinal, which embeds $(\Delta^1_2 - \mathrm{CA}) + (\mathrm{BI}) + (\Pi^1_2 - \mathrm{CA})^-$, where $(\Pi^1_2 - \mathrm{CA})^-$ is parameter free $\Pi^1_2 - \mathrm{CA}$. These theories have strength well below that of ZF, and already here interesting phenomena were discovered which were very difficult to harness proof theoretically. Writing down those results has taken a long time. Most likely the reason why an analysis has been so difficult is that our technology is not evolved enough to harness that strength. However, as long as we have not analysed proof theoretically full set theory, it cannot be ruled out that there is an inconsistency lurking somewhere.[3]

Martin-Löf said in his talk at the conference "100 years of intuitionism" at Cerisy ([24], p. 254) that we are not certain that set theory is consistent. He stressed his point using a quote by Woodin[4] He talked as well about the second failure of Hilbert's problem, which is due to technical difficulties in reaching $\Pi^1_2 - \mathrm{CA}$ and beyond[5].

Many mathematicians have experienced that sometimes when they get stuck with proving a theorem the underlying reason is that the theorem is actually false. This psychological argument does not prove anything, especially, since when getting mathematically stuck, often all that is needed is a better idea in order to prove the theorem. However, it should provide at least for the highly sceptical scientist a strong motivation to continue with the proof theoretic project. Hilbert said "We need to know, we will know".[6] The future development of proof theory will hopefully decide whether set theory is consis-

[3] And even if we have, a validation argument needs to be carried out.

[4] "Just as those who study large cardinals must admit the possibility that the notions are not consistent" [43, p. 330].

[5] Martin-Löf puts $\Pi^1_2 - \mathrm{CA}$ on the other side of the "abyss", because the analysis by Rathjen only reduces it to some set theoretic ordinal notation system. Rathjen is here following a successful tradition in the Schütte school of proof theory, and the author believes that this is already the major step in constructivising this theory. The author does not see at this moment any principal reason apart from effort and time why the resulting ordinal notation system cannot be proved to be well-founded in a suitable constructive theory. However, as long as such a reduction to a fully constructive theory has not been carried out, the analysis by Rathjen remains incomplete, and one could therefore at this moment in time place $\Pi^1_2 - \mathrm{CA}$, as Martin-Löf did, on the other side of the "abyss". See however the discussion in Sect. 5 about the limits of constructivism, which indicates that it might be very difficult to carry out the necessary constructivisation.

[6] German: "Wir müssen wissen. Wir werden wissen."

tent or not.[7] Of course, even if one ever found an inconsistency, it most likely has no effect on everyday mathematics (which is often anyway on the surface carried out in naive set theory, which is inconsistent).

2 Well-foundedness of Ordinal Notation Systems

Since the work of Gentzen, the main step in proving the consistency of reference theories in proof theory is ordinal analysis; other theories are then reduced using various techniques to these reference theories.[8] Ordinal analysis amounts to showing that the consistency of a theory can be shown in $\mathrm{PRA} + \mathrm{TI}^{\mathrm{qf}}(\alpha)$. Here PRA is primitive recursive arithmetic, and $\mathrm{TI}^{\mathrm{qf}}(\alpha)$ is the principle of quantifier free transfinite induction up to α for a specific ordinal notation system. The formula $\mathrm{TI}^{\mathrm{qf}}(\alpha)$ is defined as follows: Let $\varphi(x)$ be a quantifier free formula in the language of PRA. The formula $\mathrm{Prog}(\varphi, \alpha)$, meaning φ is progressive up to α, is defined as $\forall \beta < \alpha.((\forall \gamma < \beta.\varphi(\gamma)) \to \varphi(\beta))$. Now $\mathrm{TI}^{\mathrm{qf}}(\alpha)$ is the statement that for all such quantifier-free formulae φ we have that $\mathrm{Prog}(\varphi, \alpha)$ implies $\forall \beta < \alpha.\varphi(\beta)$. We will in the following sometimes replace in notions such as $\mathrm{TI}^{\mathrm{qf}}(\alpha)$ the ordinal α by an ordinal notation system $(A, <)$. Here, an ordinal notation system $(A, <)$ is a linearly ordered set $(A, <)$, such that A is a primitive recursive subset of $\mathbb{N}$ and $< \subseteq A \times A$ is primitive recursive. So with notations such as $\mathrm{TI}^{\mathrm{qf}}(\alpha)$ we introduce as well for ordinal notation systems $(A, <)$ the notion $\mathrm{TI}^{\mathrm{qf}}(A, <)$ for which we write as well $\mathrm{TI}^{\mathrm{qf}}(A)$.

We assume that Tait's article [41], in which he argues that PRA corresponds to finitary methods, provides sufficient arguments for validating the proof principles of PRA. So in order to validate $\mathrm{PRA} + \mathrm{TI}^{\mathrm{qf}}(\alpha)$, one needs to validate the principle of $\mathrm{TI}^{\mathrm{qf}}(\alpha)$. So assume φ is progressive up to α. Since φ is quantifier free, it is decidable, and we get $\varphi(\beta) \vee \neg\varphi(\beta)$, and can argue indirectly. Assume that for $\beta_0 := \beta$ we have that $\varphi(\beta_0)$ does not hold. Then by searching through all ordinal notations and using the decidability of φ, we can find recursively an ordinal $\beta_1 < \beta$ such that $\neg\varphi(\beta_1)$ holds. Continuing we find β_2 such that $\neg\varphi(\beta_2)$ holds. By continuing his process we obtain a recursive sequence $\beta_0 > \beta_1 > \cdots$ such that $\neg\varphi(\beta_i)$ holds for all i. Note that this argument requires Markov's principle, however not as a principle of our theory, but as a metamathematical principle. Note as well that, if we have any proof of a theorem which is not correct, it must contain (unless there is a problem with PRA) a concrete quantifier free φ and a concrete $\beta < \alpha$ for which the principle of transfinite induction up to $\beta < \alpha$ is violated. From φ and β we will then obtain a concrete infinite descending sequence. So in order to validate our theory, we need to validate

[7] Of course in case of a positive answer a validation argument needs then to be carried out.

[8] Of course often consistency is shown using normalisation proofs without ordinal analysis, however, as pointed out before when quoting the referee in Subsect. 1.1, in a proof theoretic analysis a reduction to a quite different (very slim) theory is carried out whereas in normalisation proofs we usually reduce the consistency to a slight extension of the theory in question, and therefore do not gain such a deep understanding of the proof theoretically strong principles.

that there is no recursive infinite descending sequence of ordinals $< \alpha$, which we call NRDS(α).

We will look now at the steps towards validating that ϵ_0 is well-founded. First of all, we can rule out an infinite descending recursive descending sequence of natural numbers and therefore validate NRDS(ω). If we assume NRDS($A, <_A$) and NRDS($B, <_B$) for linearly ordered sets $(A, <_A)$ and $(B, <_B)$ we can validate NRDS($A \times B, <_{\text{lex}}$) where $<_{\text{lex}}$ is the lexicographic ordering on $A \times B$ w.r.t. $<_A, <_B$. For if we had an infinite descending sequence $(a_n, b_n)_{n\in\mathbb{N}}$, we immediately see that $a_0 \geq_A a_1 \geq_A a_2 \geq \cdots$. Furthermore, for every n we can find $m > n$ s.t. $a_m <_A a_n$. For as long as $a_n = a_m$ for $n < m$ we have $b_n >_B b_{n+1} >_B .. >_B b_m$. This descending recursive sequence of b_i will eventually stop, so there must be an $m > n$ s.t. $a_m <_A a_n$, which we can find recursively. By iterating this we find an infinite descending sequence $(a_{n_k})_{k\in\omega}$ in A, which does not exist. Note that the purpose of this exercise is not proving in a formal theory $\text{TI}^{\text{qf}}(A \times B)$ but that we can get a direct insight into NRDS($A \times B$) and therefore of $\text{TI}^{\text{qf}}(A \times B)$.

Up to now we were working with recursive sequences, which corresponds to quantifier free induction. Using the validation of well-foundedness of ω and of the lexicographic ordering on the products, we can validate transfinite induction up to ω^n which is provable in PRA which has proof theoretic ordinal ω^ω. In order to prove transfinite induction up to an ordinal $\alpha \geq \omega^\omega$, quantifier free induction on ω is no longer sufficient. This translates into the non-existence of descending (possibly non-recursive) sequences in α, which we call NDS(α). For instance induction over arbitrary arithmetical formulae corresponds to non-existence of arithmetically definable descending sequences in ω. Note that NDS(α) implies NRDS(α) which as stated before validates $\text{TI}^{\text{qf}}(\alpha)$.

So we will now, instead of validating NRDS(α), validate the stronger principle NDS(α), which means we leave a fully constructive approach[9]. Even if it is nonconstructive, we consider it still to be possible to carry out a validation argument based on this notion. We can in our opinion validate NDS(ω), which means we can get a direct insight that this principle is valid. Using the same argument as before we can in our opinion validate that the principle NDS is closed under forming the lexicographic ordering for the product of two orderings.

Now assume NDS($A, <_A$). Consider A_{dec}, the set of finite sequences (or lists) of elements $(a_1, \ldots, a_k)$ of A such that $a_1 >_A \cdots >_A a_k$. Let $<_{\text{lex}}$ be the lexicographic ordering on finite sequences of elements in A based on $<_A$. We validate NDS($A_{\text{dec}}, <_{\text{lex}}$). Assume a descending sequence $(a_{n,0}, a_{n,1}, \ldots, a_{n,k_n-1})_{n\in\omega}$. We immediately see that $a_{n,0}$ is defined (i.e. $k_n \geq 1$) and weakly descending, i.e. $a_{0,0} \geq_A a_{1,0} \geq_A a_{2,0} \geq_A \cdots$. Because there is no infinite descending sequence in A, this sequence must eventually become constant. Assume it is constant from $n = n_0$ onwards. Then for $n \geq n_0$ we have that $a_{n,1}$ is defined (i.e. $k_n \geq 2$) and forms a descending sequence $a_{n_0,1} \geq_A a_{n_0+1,1} \geq_A a_{n_0+2,1} \geq_A$

[9]Constructive, if one regards Markov's principle as constructive .
In fact we will need NDS($A', <_{A'}$) only for intermediate notation systems $(A', <_{A'})$ used for validating NDS(α). For the final system, NRDS(α) is all what is required, which is implied by NDS(α).

$\cdots$ in A. That sequence will eventually become constant for $n \geq n_1$ for some n_1. Therefore $a_{n,2}$ is descending for $n \geq n_1$ onwards and will become constant for $n \geq n_2$ for some n_2. By continuing this process we obtain a sequence of natural numbers $(n_i)_{i\in\omega}$ and have $a_{n_0,0} = a_{n_1,0} >_A a_{n_1,1} = a_{n_2,1} >_A a_{n_2,2} >_A \cdots$. So we obtain an infinite descending sequence $a_{n_0,0} >_A a_{n_1,1} >_A \cdots$ in A which does not exist, and have therefore shown that there is no infinite descending sequence in $(A_{\mathrm{dec}}, <_{\mathrm{lex}})$. Note that we cannot determine $n_0, n_1, \cdots$, so NRDS(A) is not sufficient to carry out this argument.

This argument validates transfinite induction on $(A_{\mathrm{dec}}, <_{\mathrm{lex}})$. Ordering on ordinals in Cantor Normal Form (CNF) $\alpha = \omega^{\alpha_1} n_1 + \cdots \omega^{\alpha_k} n_k$ is the same as the double lexicographic ordering on $((\alpha_1, n_1), \ldots, (\alpha_k, n_k))$. Let $(A, <)$ be an ordinal notation system. Let CNF(A) be the set of terms obtained by applying once CNF to elements in A, ordered correspondingly. CNF(A) is isomorphic to a subset of $((A \times (\omega \setminus 0), <_{\mathrm{lex}})_{\mathrm{dec}}, <_{\mathrm{lex}})$[10] which in turn is isomorphic to $((A \times \omega, <_{\mathrm{lex}})_{\mathrm{dec}}, <_{\mathrm{lex}})$. The order type of CNF($A$) is ω^α, if the order type of A is α. This means that, if we have validated NDS(α), we have validated NDS(ω^α).

Therefore we can validate NDS(ω_n) and therefore at least $\mathrm{TI}^{\mathrm{qf}}(\omega_n)$ where $\omega_0 = \omega$, $\omega_{n+1} = \omega^{\omega_n}$. Since $\epsilon_0 = \sup_{n\in\omega} \omega_n$ we have validated quantifier free transfinite induction up to all ordinals less than ϵ_0.

Gentzen showed that PRA+$\mathrm{TI}^{\mathrm{qf}}(\epsilon_0)$ proves the consistency of PA, which was considered as a proof of the consistency of PA. The belief that this proof shows the consistency of PA (in an absolute way) must be based on some argument which validates PRA + $\mathrm{TI}^{\mathrm{qf}}(\epsilon_0)$, and we have given one such argument. The above argument has shown the validity of the consistency of PA. Therefore it follows, for instance, that, if we have shown in PA Fermat's last theorem, then there can be no counter example.

In our articles [36, 37] we extended this approach to ordinal notation systems from below. Up to the strength of $(\Pi^1_1 - \mathrm{CA})_0$ we were able to give arguments, which we regard as a validation of transfinite quantifier-free induction up to those ordinals. When reaching higher ordinals, the direct insight into the well-foundedness rests necessarily upon principles of increasing proof theoretic strength. Note that according to the results of reverse mathematics, most real mathematical theorems can be shown in $(\Pi^1_1 - \mathrm{CA})_0$, so most of mathematics can be validated by pure ordinal analysis. Beyond that strength, we could develop ordinal notation systems from below, but could only give a formal well-foundedness proof, which then needs to be carried out in another theory of at least equal strength. It is no accident that this happens when moving from $(\Pi^1_1 - \mathrm{CA})_0$ to $\Pi^1_1 - \mathrm{CA}$, since the argument is based on the concept of well-foundedness, which is a Π^1_1-concept, and one needs in some form a principle, which goes beyond Π^1_1, in order to validate $\Pi^1_1 - \mathrm{CA}$.

[10] Those sequences $((a_1, n_1), \ldots, (a_k, n_k))$ s.t.$a_1 > \cdots > a_k$.

3 Martin-Löf Type Theory

With increasing strength, ordinal notation systems for describing the proof theoretic ordinal of theories become increasingly complicated. Therefore, the complexity of the well-foundedness proofs for these ordinal notation systems increases as well. Correspondingly, it becomes increasingly difficult, if possible at all, to validate the well-foundedness of the ordinal notation system directly. A solution for this problem is to make a second step and prove the well-foundedness of the ordinal notation system in a second theory for which one can carry out a validation argument more directly. Hilbert wanted originally to validate theories involving the infinite by reducing them to finitary methods. A suitable generalisation of finitary methods are constructive theories, in which the elements of sets are still finite objects, or terms. In order to deal with function spaces, we need reduction rules for terms, for instance $n + \mathrm{S}(m)$ reduces to $\mathrm{S}(n + m)$. This allows to determine elements of function types as terms which applied to elements of the argument type are elements of the result type, or reduce to such an element. So infinite objects (full functions) are replaced by finite objects (programs or terms).

The addition of recursive functions as finitary objects was the motivation of Gödel in his Dialectica paper ([13]), where he writes (p. 282, translation p. 245 of [11]): "It is the second requirement that must be dropped. This fact has hitherto been taken into account by our adjoining to finitary mathematics parts of intuitionistic logic and the theory of ordinals. In what follows we shall show that, for the consistency proof of number theory, we can use, instead, the notion of computable function of finite type on the natural numbers and certain elementary principles of construction for such functions."[11]

Gödel's Dialectica interpretation was still referring to classical logic, and is usually used mainly as a proof theoretical tool rather than being considered as an approach to obtaining a foundation of mathematics. A more radical approach was taken in Martin-Löf's type theory (MLTT)[12]. MLTT is, as Martin-Löf phrased it once to the author (we unfortunately do not remember the precise wording), the most serious attempt to develop a theory such that we have an insight that all judgements are valid. Those not familiar with MLTT are often perplexed by the large number of its rules. The reason for having such a large number of rules is that this theory is not defined so that it has a shortest

[11] "Es ist die zweite Forderung, welche fallen gelassen werden muss. Dieser Tatsache wurde bisher dadurch Rechnung getragen, dass man Teile der intuitionistischen Logik und Ordinalzahltheorie zur finiten Mathematik adjungierte. Im folgenden wird gezeigt, dass man statt dessen für den Widerspruchsfreiheitsbeweis der Zahlentheorie auch den Begriff der berechnenbaren Funktion endlichen Types über den natürlichen Zahlen und gewisse sehr elementare Konstruktionsprinzipien für solche Funktionen verwenden kann."

[12] The standard reference is Martin-Löf's book [20]. The article [28] contains a good concise summary of the rules of MLTT (starting p. 162), however the rules for ω and Ω, which make it a partial type theory, the topic of that article, need to be omitted. Another listing can be found in the author's article [35], where everything was made precise in order to be able to carry out a proof theoretic analysis. Arne Ranta's book [29] contains a nice introduction to MLTT. Nordström et al.'s book [26] is an excellent reference for MLTT, and there is the more recent and more concise handbook version [25].

description. Instead it is designed so that we can get an insight into the validity of all provable judgements.

In MLTT we have non-dependent judgements of the form

- $a : A$ for a is of type A,
- $a = b : A$ for a, b are equal elements of type A,
- $A : \mathrm{Set}$ for A is a set,
- $A = B : \mathrm{Set}$ for A, B are equal sets.

Dependent judgements have the form $x_1 : A_1, \ldots, x_n : A_n \Rightarrow \theta$ where θ is a non-dependent judgement, with free variables in $x_1, \ldots, x_n$.

We have as rules

- structural rules (rules for dealing with contexts, assumptions, and the definitional equalities $a = b : A$ and $A = B : \mathrm{Set}$);
- formation rules (which introduce sets, e.g. conclude $\mathbb{N} : \mathrm{Set}$);
- introduction rules (which introduce a canonical element, an element starting with a constructor, e.g. for $\mathbb{N}$ derive $0 : \mathbb{N}$ and from $a : \mathbb{N}$ derive $\mathrm{S}(a) : \mathbb{N}$);
- elimination rules, e.g. higher type primitive recursion in case of $\mathbb{N}$,
- equality rules (e.g. deriving that if $t(x)$ is defined by higher type primitive recursion into type $B(x)$, with base case $a : B(0)$, that $t(0) = a : B(0)$);
- and equality versions of the formation, introduction and elimination rules (e.g. deriving $\mathrm{S}(a) = \mathrm{S}(a') : \mathbb{N}$ from $a = a' : \mathbb{N}$).

The validation argument for MLTT is done via meaning explanations.[13] In meaning explanations, one determines the meaning of each judgement. Then one validates for each rule that, if the premises are valid w.r.t. meaning explanations, so is the conclusion. Therefore all judgements provable are valid.

Elements of sets can be canonical elements, which are formed by the introduction rules. For instance, $\mathrm{S}(2+2)$ is a canonical element of $\mathbb{N}$. Non-canonical elements are considered by Martin-Löf (see, e.g., [20]) as programs, which evaluate to a canonical element. Canonical elements are special cases of non-canonical elements, which as programs evaluate to themselves. Martin-Löf (private communication) considers the concept of a program, for which we have a direct insight how it operates, as crucial for understanding his meaning explanations.

[13] We could not find a definite and complete reference to meaning explanations. Martin-Löf's articles and book [20, 21, 22, 23] introduce meaning explanations when discussing the rules of type theory. Tasistro's PhD thesis [42] describes meaning explanations if one uses explicit substitutions (see as well a short reference in the more accessible article [12]). The author has in [39] given an account of his understanding of meaning explanation with a variation in order to accommodate coalgebraic data types defined by their elimination rules.

The meaning of $A : \mathrm{Set}$ is given by determining what its canonical elements are and when two canonical elements are equal. The meaning of $a : A$ is that a is a non-canonical element of A. The meaning of the judgement $a = a' : A$ is that a, a' are equal elements of A, which means that they evaluate to equal canonical elements of A.

In case of $\mathbb{N}$ we have that 0 is a canonical element, and, if n is an element of $\mathbb{N}$, then $\mathrm{S}(n)$ is a canonical element of it. 0 is equal to 0, and if n, m are two equal elements of $\mathbb{N}$, then $\mathrm{S}(n)$ and $\mathrm{S}(m)$ are equal canonical elements of it.

The meaning of the judgement $A = B : \mathrm{Set}$ is that A and B are equal sets which means that canonical elements of A are canonical elements of B and vice versa, and equal canonical elements of A are equal canonical elements of B and vice versa.

For determining the meaning of dependent judgements, we introduce abbreviations $\vec{x}$ for $x_1, \ldots, x_n$, similar for $\vec{a}$, $\vec{a}'$ (referring to a'_i), and $\vec{x}_k$ for $x_1, \ldots, x_k$, similar for $\vec{a}_k, \vec{a}'_k$. A dependent judgement

$$x_1 : A_1, x_2 : A_2(x_1), \ldots, x_n : A_n(\vec{x}_{n-1}) \Rightarrow \theta(\vec{x})$$

is valid if for every choice of elements

$$a_1 : A_1, a_2 : A_2(a_1), \ldots, a_n : A_n(\vec{a}_{n-1})$$

the judgement $\theta(\vec{a})$ is valid. One needs as well that for equal elements

$$a_1 = a'_1 : A_1, a_2 = a'_2 : A_2(a_1), \ldots, a_n = a'_n : A_n(\vec{a}_{n-1})$$

the equality judgements in the conclusion holds: If $\theta = (A : \mathrm{Set})$ we require that $A(\vec{a}) = A(\vec{a}') : \mathrm{Set}$ holds, in case $\theta = (a : A)$ we require that $a(\vec{a}) = a(\vec{a}') : A(\vec{a})$ holds. Judgement $A = B : \mathrm{Set}$ presupposes $A : \mathrm{Set}$, $B : \mathrm{Set}$, judgement $a : A$ presupposes $A : \mathrm{Set}$, judgement $a = b : A$ presuppose $a : A$, $b : A$. The judgement

$$x_1 : A_1, \ldots, x_n : A_n(\vec{x}_{n-1}) \Rightarrow \theta(\vec{x})$$

presupposes $A_1 : \mathrm{Set}$, $x_1 : A_1 \Rightarrow A_2(x_1) : \mathrm{Set}$, etc, and as well

$$x_1 : A_1, \ldots, x_n : A_n(\vec{x}_{n-1}) \Rightarrow \theta'(\vec{x})$$

for any presupposition $\theta'(\vec{x})$ of $\theta(\vec{x})$.

Adding the meaning of the presuppositions of judgements (applied transitively) to the meaning of a judgement gives the full meaning of the judgement.

Now one can easily validate structural rules, formation rules, introduction rules, and their equality versions. Elimination rules are more difficult to validate (and that's where an increasingly high level of trust is required). In case of $\mathbb{N}$, in the simple case where we derive $x : \mathbb{N} \Rightarrow t(x) : B(x)$ by primitive recursion, we validate that $t(0) : B(0)$ and if we have $x : \mathbb{N}$ and $t(x) : B(x)$ are valid, so is $t(\mathrm{S}(x)) : B(\mathrm{S}(x))$. Now one sees that for each element of a of $\mathbb{N}$ as given by the meaning explanations $t(a) : B(a)$. This holds first for canonical elements, by going through what we said constitutes a canonical element of $\mathbb{N}$, and checking

for each canonical element a that $t(a) : B(a)$ is validated. For non-canonical elements, the reduction of $t(a)$ is given by first reducing a to canonical form 0 or $\mathrm{S}(a')$, and then applying the reductions corresponding to the base case or induction step. Therefore the rules are validated as well for non-canonical elements.

The key principle one needs to trust is the correctness of the elimination rules for the inductively defined sets $\mathbb{N}$, W-type, and universes. We cannot get around the fact that we cannot prove the consistency of MLTT, so when moving to proof theoretically stronger principles, one needs to trust the validity of the rules for proof theoretically stronger sets. We cannot avoid this, but the author believes that one can trust in the principles involved.

3.1 Induction-Recursion and the Mahlo Universe

The validation of principles works well for concrete inductive-recursively defined sets, as long as we do not make use of the full logical framework, which allows to have $A : \mathrm{Set}$ or even higher types in the context.[14] Therefore, one can validate Palmgren's superuniverse ([27], Sect. 3), but not Palmgren's higher order universes ([27], Sect. 5) or the external Mahlo universe ([4], Sect. 6.3), which reaches at least the strength of KPM ([4], Sect. 6.4). The strength of Palmgren's superuniverse is not known ([30, 31] analyse only the metapredicative version without the W-type), but substantially exceeds that of Martin-Löf type theory with W and one universe.[15] The latter theory was analysed by Rathjen, Griffor, Setzer [34, 14, 35], and has strength slightly bigger than Kripke Platek set theory with one recursively inaccessible, KPI.

For the Mahlo universe we have given meaning explanations in our article [38] (not yet published). However, we cannot say that the validity of its rules are as fully convincing as they are for inductive-recursive definitions without use of the full logical framework.

[14] When introducing his version of meaning explanations, the author usually avoids the logical framework. The reason is that he has not yet found an account of meaning explanations of the logical framework, which does not consider Set as a Russell style subuniverse of Type, and which he considers as fully satisfactory. If Set is treated as a universe, one adds considerable proof theoretic strength. Especially, with the rules for inductive-recursive definitions Set is closed under the introduction rules of (a Russell style variant of) the internal Mahlo universe. In the community of MLTT, inductive-recursive definitions is often considered as the limit of what can be at the moment justified without making use of the Mahlo universe principle. Martin-Löf has given presentations about how to treat the logical framework without adding additional strength, however we could not find yet a written account of it needed in order to judge it completely.

[15] It is easy to conjecture the precise strength, and it would not be difficult albeit time consuming to analyse the full version of it.

4 Feferman's System of Explicit Mathematics and the Extended Predicative Mahlo Universe

In [16] Kahle and the author have published an extended predicative version of the Mahlo universe. This version is developed in Feferman's system of explicit mathematics [5, 6]. It uses the fact that in Feferman's system one has access to the collection of all terms, and therefore can form for every term a subuniverse of the Mahlo universe which is relatively closed under this term considered as a partial function. In MLTT all objects have a type and are therefore total. Therefore in MLTT we do not have access to the collection of all terms, which in general are only partial objects.

We regard this version ([16]) as being predicative (in an extended sense) and believe that this theory can be validated. Feferman's theory has been developed in second order logic[16], and optimised towards a short and concise theory. While this makes metamathematical investigations easy and makes it easily accessible to non-specialists, it causes problems when validating the provable statements[17]. It seems however that this is not a principal problem. It should be possible to present Feferman's theories in a style which is very close to that of Martin-Löf type theory, and develop meaning explanations. This way hopefully one could validate the extended predicative Mahlo universe in the sense of this article.

With [16] we have not reached the limit of this methodology. We have developed draft versions which reach at least the strength of Kripke Platek set theory extended by Π_3-reflection, and it is likely that we can go far beyond with that strength.

5 The Limit of Constructivism

In [32, Sect. 6] Rathjen introduces assumptions (A0) - (A3) about possible extensions MLTT$^+$ of Martin-Löf Type Theory, of which the most important one is assumption (A3):

(A3) Every inductive definition $\Phi : \mathrm{Pow}(\mathbb{N}) \to \mathrm{Pow}(\mathbb{N})$ for generating the elements of a type A in MLTT$^+$ and its pertinent decoding function are definable by set-theoretic Σ-formulae. These formulae may contain further sets as parameters, where these sets correspond to previously defined types.

He shows (Theorem 6.1) that under these assumptions a set M such that $M \prec_1 \mathrm{V}$ is a model of MLTT$^+$. Here $M \prec_1 \mathrm{V}$ means that M is a Σ_1-elementary substructure of V, where V is the set theoretic universe. This determines a limit to a constructive program based on MLTT.

[16] Not much of second order logic is actually used, its use is mainly for convenience rather than need.

[17] We note that this is the opinion of the second author of [16] only, who is the author of the current article.

In his argument, Rathjen already admits that due to the acceptance of the Mahlo universe as an acceptable extension of MLTT, a more strict assumption had to be abandoned, namely that sets are introduced by monotone inductive definitions. This already indicates that it is very difficult to determine an upper bound for a constructive program. While it may be difficult to go beyond principle (A3), we believe that this is only a temporary limitation – it is likely that new constructive principles will emerge, which will be considered as acceptable but go beyond this principle. However, drawing this line is of great benefit, since it determines the requirements a new principle needs to fulfil in order to go beyond that limit.

Acknowledgements The author wants to thank the anonymous referees for extraordinarily detailed refereeing and many very valuable comments; Fredrik Nordvall Forsberg and Håkon Gylterud for careful proof reading; and Reinhard Kahle for his encouragement to writing such a rather philosophical article and for his patience while waiting for the completion of this article. Research for this article was supported by EPSRC grant EP/G033374/1.

References

[1] T. Arai. Proof theory of theories of ordinals III: Π_1 collection. Unpublished Notes, 1997.

[2] T. Arai. A sneak preview of proof theory of ordinals. arXiv:1102.0596v1 [math.LO], http://arxiv.org/abs/1102.0596, 2011.

[3] P. Dybjer. Program testing and the meaning explanations of intuitionistic type theory. In *Epistemology versus Ontology*, pages 215–241. Springer, 2012.

[4] P. Dybjer and A. Setzer. Induction-recursion and initial algebras. *Annals of Pure and Applied Logic*, 124:1 – 47, 2003.

[5] S. Feferman. A language and axioms for explicit mathematics. In J. Crossley, editor, *Algebra and Logic*, volume 450 of *Lecture Notes in Mathematics*, pages 87–139. Springer Berlin / Heidelberg, 1975. 10.1007/BFb0062852.

[6] S. Feferman. Constructive theories of functions and classes. In D. D. Maurice Boffa and K. McAloon, editors, *Logic Colloquium '78 Proceedings of the colloquium held in Mons*, volume 97 of *Studies in Logic and the Foundations of Mathematics*, pages 159 – 224. North-Holland, Amsterdam, New York, Oxford, 1979.

[7] S. Feferman. Hilbert's program relativized: Proof-theoretical and foundational reductions. *The Journal of Symbolic Logic*, 53(2):pp. 364–384, 1988.

[8] S. Feferman. What rests on what? The proof-theoretic analysis of mathematics. In J. Czermak, editor, *Philosophy of Mathematics. Proceedings of the Fifteenth International Wittgenstein-Symposium, Part 1*, pages 147–171, Vienna, 1993. Hölder-Pichler-Tempsky. Reprinted in Solomon Feferman: In the light of Logic. Oxford University Press, 1998, Ch. 10, 187 – 208.

[9] S. Feferman. Why a little bit goes a long way: Logical foundations of scientifically applicable mathematics. *PSA*, 2:442 – 455, 1993. Reprinted in Feferman: In the light of logic. Oxford University Press, 1998, Ch. 14, pp. 284 – 298.

[10] S. Feferman. Does reductive proof theory have a viable rationale? *Erkenntnis*, 53(1 – 2):63 – 96, September 2000.

[11] S. Feferman, J. W. Dawson Jr, S. C. Kleene, G. H. Moore, R. M. Solovay, and J. van Heijenoort, editors. *Kurt Gödel. Collected Works: Volume II: Publications 1938-1974*. Oxford University Press, 1990.

[12] D. Fridlender. A proof-irrelevant model of Martin-Löf's logical framework. *Mathematical Structures in Computer Science*, 12:771–795, 12 2002.

[13] K. Gödel. Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes. *Dialectica*, 12(3 – 4):280 – 287, December 1958. Translation in [11], pp. 240 – 251.

[14] E. Griffor and M. Rathjen. The strength of some Martin-Löf type theories. *Archive for Mathematical Logic*, 33(5):347, 1994.

[15] H. Hesse. *Das Glasperlenspiel. Versuch einer Lebensbeschreibung des Magister Ludi Josef Knecht samt Knechts hinterlassenen Schriften. (English: The glass bead game)*. Fretz & Wasmuth, Zürich, 1943.

[16] R. Kahle and A. Setzer. An extended predicative definition of the Mahlo universe. In R. Schindler, editor, *Ways of Proof Theory*, Ontos Series in Mathematical Logic, pages 309 – 334. Ontos Verlag, 2010.

[17] G. Kreisel. A variant of Hilbert's theory of the foundations of arithmetic. *Br J Philos Sci*, IV(14):107 – 129, 1953.

[18] G. Kreisel. A survey of proof theory. *The Journal of Symbolic Logic*, 33(3):pp. 321–388, 1968.

[19] G. Kreisel. Hilbert's programme. In P. Benacerraf and H. Putnam, editors, *Philosophy of mathematics: selected readings*, pages 207 – 238. Cambridge University Press, 2nd edition, 1983. Revised version of article in Dialectica 12(3 -4), pp. 346 – 372, 1958.

[20] P. Martin-Löf. *Intuitionistic type theory*. Bibliopolis, Naples, 1984.

[21] P. Martin-Löf. Truth of a proposition, evidence of a judgement, validity of a proof. *Synthese*, 73:407 – 420, 1987.

[22] P. Martin-Löf. On the meaning of the logical constants and the justification of the logical laws. *Nordic Journal of Philosophical Logic*, 1(1):11 – 60, May 1996. Short course given at the meeting Teoria della Dimostrazione e Filosofia della Logica, organized in Siena, 6-9 April 1983, by the Scuola di Specializzazione in Logica Matematica of the Università degli Studi di Siena. Available from http://www.hf.uio.no/filosofi/njpl/vol1no1/meaning/meaning.html.

[23] P. Martin-Löf. An intuitionistic theory of types. In G. Sambin and J. Smith, editors, *Twenty-Five Years of Constructive Type Theory*, pages 127 – 172, Oxford, 1998. Oxford University Press. Reprinted version of an unpublished report from 1972.

[24] P. Martin-Löf. The Hilbert-Brouwer controversy resolved? In M. Atten, P. Boldini, M. Bourdeau, and G. Heinzmann, editors, *One Hundred Years of Intuitionism (1907 – 2007)*, Publications des Archives Henri Poincaré, Publications of the Henri Poincaré Archives, pages 243–256. Birkhäuser Basel, 2008.

[25] B. Nordstrom, K. Petersson, and J. Smith. Martin-löf's type theory. In S. Abramsky, D. M. Gabbay, and T. Maibaum, editors, *Handbook of Logic in Computer Science: Logic and algebraic methods, Volume 5*, pages 1 – 38. Oxford University Press, USA, 2001.

[26] B. Nordström, K. Petersson, and J. M. Smith. *Programming in Martin-Löf's type theory. An introduction.* Oxford University Press, 1990. Book out of print. Online version available via http://www.cs.chalmers.se/Cs/Research/Logic/book/.

[27] E. Palmgren. On universes in type theory. In G. Sambin and J. Smith, editors, *Twenty five years of constructive type theory*, pages 191 – 204, Oxford, 1998. Oxford University Press.

[28] E. Palmgren and V. Stoltenberg-Hansen. Domain interpretations of Martin-Löf's partial type theory. *Annals of Pure and Applied Logic*, 48:135 – 196, 1990.

[29] A. Ranta. *Type-theoretical grammar.* Oxford University Press, 1995.

[30] M. Rathjen. The strength of Martin-Löf type theory with a superuniverse. Part I. *Archive for Mathematical Logic*, 39(1):1 – 39, January 2000.

[31] M. Rathjen. The strength of Martin-Löf type theory with a superuniverse, part II. *Archive for Mathematical Logic*, 40(3):207–233, 2001.

[32] M. Rathjen. The constructive Hilbert program and the limits of Martin-Löf type theory. *Synthese*, 147:81 – 120, 2005.

[33] M. Rathjen. An ordinal analysis of parameter free Π^1_2-comprehension. *Arch. Math. Log.*, 44(3):263 – 362, April 2005.

[34] A. Setzer. *Proof theoretical strength of Martin-Löf Type Theory with W-type and one universe.* PhD thesis, Universität München, available via http://www.cs.swan.ac.uk/∼csetzer, 1993.

[35] A. Setzer. Well-ordering proofs for Martin-Löf type theory. *Annals of Pure and Applied Logic*, 92:113 – 159, 1998.

[36] A. Setzer. Ordinal systems. In S. B. Cooper and J. Truss, editors, *Sets and Proofs*, pages 301 – 331, Cambridge, 1999. Cambridge University Press.

[37] A. Setzer. Ordinal systems part 2: One inaccessible. In S. Buss, P. Hajek, and P. Pudlak, editors, *Logic Colloquium '98*, ASL Lecture Notes in Logic 13, pages 426 – 448, Massachusetts, 2000. Peters Ltd.

[38] A. Setzer. Universes in type theory Part II – Autonomous Mahlo. Submitted to Annals of Pure and Applied Logic. Available from http://www.cs.swan.ac.uk/∼csetzer/articles/modelautomahlomain.pdf, 2011.

[39] A. Setzer. Coalgebras as types determined by their elimination rules. In P. Dybjer, S. Lindström, E. Palmgren, and G. Sundholm, editors, *Epistemology versus Ontology*, volume 27 of *Logic, Epistemology, and the Unity of Science*, pages 351–369. Springer Netherlands, 2012. 10.1007/978-94-007-4435-6_16.

[40] Stanford Encyclopedia of Philosophy. Hilbert's program. Available from http://plato.stanford.edu/entries/hilbert-program/#1, 2003.

[41] W. W. Tait. Finitism. *The Journal of Philosophy*, 78(9):524–546, Sep. 1981.

[42] A. Tasistro. *Substitution, record types and subtyping in type theory, with applications to the theory of programming.* PhD thesis, Department of Computing Science, University of Gothenburg, Gothenburg, Sweden, 1997.

[43] W. H. Woodin. The tower of Hanoi. In H. H. G. Dales and G. Oliveri, editors, *Truth in mathematics*, pages 329–351. Oxford University Press, 1998.

[44] R. Zach. Hilbert's program then and now. In D. Jacquette, editor, *Philosophy of Logic*, pages 411 – 447. North-Holland, Amsterdam, 2007.